Explaining Variations in Hospital Death Rates

Randomness, Severity of Illness, Quality of Care

Rolla Edward Park, Robert H. Brook,
Jacqueline Kosecoff, Joan Keesey,
Lisa V. Rubenstein, Emmett B. Keeler,
Katherine L. Kahn, William H. Rogers,
Mark R. Chassin

RAND

The research described in this report was supported by the Health Care Financing Administration, U.S. Department of Health and Human Services, Cooperative Agreement 18-C-98684/9-02.

Library of Congress Cataloging in Publication Data

Explaining variations in hospital death rates : randomness, severity of illness, quality of care / Rolla Edward Park ... [et al.].

p. cm.

"Supported by the Health Care Financing Administration."

"R-3887-HCFA."

ISBN 0-8330-1098-0

1. Hospital patients—United States—Mortality—Statistics. 2. Congestive heart failure—Patients—United States—Mortality—Statistics. 3. Heart—Enfarction—Patients—United States—Mortality—Statistics. I. Park, Rolla Edward. II. RAND Corporation. III. United States. Health Care Financing Administration.

[DNLM: 1. Health Insurance for Aged and Disabled, Title 18. 2. Hospitalization. 3. Mortality—United States. 4. Quality of Health Care—United States. 5. Severity of Illness Index. WA 900 AA1 E9]

RA981.A2E965 1990

362.1'96129025'00973—dc20

DNLM/DLC

for Library of Congress 90-9213

CIP

Published 1991 by RAND
1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138

R-3887-HCFA

Explaining Variations in Hospital Death Rates

Randomness, Severity of Illness, Quality of Care

Rolla Edward Park, Robert H. Brook,
Jacqueline Kosecoff, Joan Keesey,
Lisa V. Rubenstein, Emmett B. Keeler,
Katherine L. Kahn, William H. Rogers,
Mark R. Chassin

Supported by the
Health Care Financing Administration,
U.S. Department of Health and Human Services

RAND

PREFACE

Sherry A. Terrell, PhD
HCFA, ORD,OR,DRES,NISB
2-B-14 Dakm Bldg.
6325 Security Blvd.
Baltimore, MD 21207

Each year since 1986, the Health Care Financing Administration has made public analyses of individual hospital death rates for Medicare patients during the previous year (or two or three). Death rates vary markedly, even for patients hospitalized for the same reasons. Underlying the public release of these data is the idea that some of the variation must reflect differences in the quality of care provided in different hospitals. But some must also reflect different severity of illness for patients treated in different hospitals, and some may result from random or unmeasurable factors over which hospitals have no control.

In this report, an attempt is made to sort out the relative importance of quality of care, severity of illness, and unmeasured factors or selection effects as determinants of hospital death rates. Medicare patients with congestive heart failure and acute myocardial infarction are studied. These conditions together accounted for nearly 18 percent of all Medicare hospital deaths during 1984.

The body of the report is based on (but in some places is more detailed than) an article published in *The Journal of the American Medical Association,* Vol. 264, No. 4, July 25, 1990, copyright American Medical Association and reused by permission. The extensive appendixes include additional material that supports or expands on statements made in the body of the report.

SUMMARY

Hospital death rates vary markedly, even for the same disease. We studied a representative sample of 1126 congestive heart failure patients and 1150 acute myocardial infarction patients in hospitals with unexpectedly high disease-specific death rates ("targeted" hospitals) compared with all other ("untargeted") hospitals in four populous states (California, Illinois, Minnesota, New York), using both inpatient deaths and deaths within 30 days of admission. Death rates in targeted hospitals were 5.0 to 10.9 higher per 100 admissions than in untargeted hospitals. However, 56 to 82 percent of the excess could result from random binomial variation, even if all hospitals provided the same quality of care to the same age/sex/race mix of patients. We measured severity of illness and quality of care using detailed medical records abstracts; at the individual patient level, higher severity and lower quality were both associated with higher probability of death. However, we found only small and insignificant differences in quality between targeted and untargeted hospitals; even at a 95 percent confidence bound on the estimated difference in quality, quality differences could explain only 0.3 or fewer of the excess deaths per 100 admissions in targeted hospitals. Severity differences, too, were also small for hospitals treating congestive heart failure patients. For myocardial infarction patients, however, severity differences explained up to 2.8 excess deaths per 100 admissions in targeted hospitals. There is some evidence that targeting hospitals with consistently high death rates over periods longer than one year may better identify potential quality problems.

ACKNOWLEDGMENTS

We appreciate the good work of Caren Kamberg and Barbara Genovese, who were responsible for most peer review organization contact and hardcopy data management for this study. We also appreciate helpful comments on drafts by Lincoln Moses of Stanford University; two anonymous referees for *The Journal of the American Medical Association*; and Paul Eggers, Marian Gornick, Steve Jencks, and Henry Krakauer of the Health Care Financing Administration.

CONTENTS

FIGURES

TABLES

I. INTRODUCTION

It would be convenient if hospitals with higher death rates, identified using easily collected administrative data (age, sex, previous hospitalization, and diagnosis), turned out to be providing lower quality of care. It is easy to use administrative data to identify high death rate hospitals. If a high death rate were a marker for bad care, then health care consumers would know to avoid those hospitals, and professional organizations and the hospitals themselves could work to correct the quality problems.

Apparently in the hope or belief that high death rates and low quality of care are associated, the Health Care Financing Administration (HCFA) has, annually since 1986, released increasingly sophisticated analyses of hospital death rates at individual hospitals for Medicare patients (Brinkley, 1986; Bowen and Roper, 1987, 1988; Sullivan and Hays, 1989). Release of the analyses has been criticized (Greenfield et al., 1988; Blumberg, 1987; Wagner, Knaus, and Draper, 1986), but HCFA has taken many of the criticisms into account. Even HCFA's critics seem to share the hope that sufficiently sophisticated analyses will succeed in targeting hospitals that provide substandard care.

Although variations in hospital death rates have been studied for a long time, there are few such studies, and until the release of data by HCFA, they have not been performed to identify individual hospitals as possibly providing poor quality of care (Fink, Yano, and Brook, 1989). Death rates have been shown to vary by specific hospital characteristics (Flood, Scott, and Ewy, 1984; Flood and Scott, 1987) and by experience (i.e., volume) (Hannan et al., 1989; Luft, Bunker, and Enthoven, 1979; Riley and Lubitz, 1985), but we do not know a lot about how much of the variation results from differences in severity of illness or quality of care, and how much from random or selection effects.

Some evidence has accumulated, from studies in limited numbers of hospitals, that some of the differences in death rates among hospitals may be due to differences in severity of illness or level of comorbidity. This was true for patients with pneumonia, myocardial infarction, or stroke in a single large hospital chain (Dubois et al., 1987), for patients with cancer in seven hospitals (Greenfield et al., 1988), and for patients in nine pediatric intensive care units (Pollack et al., 1987).

One recent study of four common medical conditions in a Medicare population found that chance variation could account for a major part

of the differences in hospital death rates, but that severity measures based on data obtained from a medical record review also helped to explain the differences (Jencks et al., 1988). Another study of five common conditions in 13 hospitals found that severity measured from the medical record added substantially to the explanatory power of HCFA's 1988 model and reduced instances of higher than expected mortality to chance levels (Green et al., 1990).

Only one study has shown some connection between high death rates and quality of care. Using implicit peer review of quality of care in a single hospital chain, that study showed that pneumonia, stroke, or myocardial infarction patients were twice as likely to suffer a possibly preventable death in high-death outlier hospitals than were patients in hospitals that were not statistical outliers (Dubois et al., 1987).

We previously found for all U.S. acute care hospitals that age-sex-race-disease-specific death rates were significantly different (both clinically and statistically) by hospital for 22 out of 48 specific conditions or diagnoses (Chassin et al., 1989). Because our previous study used data from hospital claims only, we were unable to address the question of how to explain the systematic variation. The present study attempts to shed some additional light on the relationships among hospital death rates, severity of illness, and quality of care by using data from medical records. We chose two medical conditions—congestive heart failure (CHF) and acute myocardial infarction (AMI)—for more detailed clinical investigation because their death rates varied significantly by hospital and they accounted for 7.5 percent of all Medicare admissions and 17.5 percent of all Medicare hospital deaths.

Our primary objectives for this study were to determine in a representative sample of acute care hospitals (1) whether hospitals with high age-sex-race-disease-specific death rates provide lower quality of care or treat more severely ill patients than do hospitals with lower death rates, and (2) how the probability of death at the patient level is related to severity of illness and quality of care. Answers to these questions could result in better policy decisions regarding whether the public identification of poor quality hospitals requires collecting more data (e.g., severity of illness at time of admission) than that available on a discharge abstract.

II. METHODS

ADMINISTRATIVE DATA

We obtained information on all hospital stays for Medicare beneficiaries from HCFA's Bill Record File for all admissions occurring between October 1, 1983, and September 30, 1984. To make the data as comparable as possible across hospitals, we (1) excluded all Medicare beneficiaries under the age of 65 (those eligible to receive Medicare benefits solely because of various disabilities, including chronic renal disease); (2) excluded data from long term care hospitals, psychiatric facilities, hospices, and rehabilitation hospitals; (3) excluded interim bills; (4) edited the data to include only one complete record for each hospital stay; and (5) counted transfers from one acute care hospital to another as live discharges from the first hospital and separate admissions to the second.

We obtained additional information on hospital characteristics from HCFA's Provider of Service File, and information on out of hospital deaths from HCFA's Health Insurance Master File. We defined congestive heart failure as DRG (diagnosis related group) 127 with a principal diagnosis of ICD-9 codes 398.91, 402.11, 402.91, 428.0, 428.1, 428.9, or 785.51, and acute myocardial infarction as DRGs 121, 122, 123, and 115 with a principal diagnosis of ICD-9 codes 410.0 through 410.9.

TARGETING

For each hospital in the administrative data base, we calculated d, the death rate it would have experienced if its congestive heart failure or acute myocardial infarction patients had died at nationwide average rates for each condition for each of 20 age-sex-race cells. We then calculated the binomial probability that a hospital whose n patients each had a true probability of dying d would have as many deaths m as it actually did $p(d,n,m)$. Hospitals with less than a 0.05 probability of having as many deaths as they did, $p(d,n,m) < 0.05$, were called targeted (high death rates); all others were untargeted. In simple terms, we targeted using a one-sided test at the 0.05 level. We targeted separately for each condition (congestive heart failure and myocardial infarction), and separately for inpatient deaths and for deaths within 30 days of admission.

We compared our targeting with HCFA's targeting in 1986 and 1987 for severe chronic and severe acute heart disease—conditions that HCFA defines more inclusively than our CHF and AMI. Also, HCFA uses different targeting methods than we do. To make comparison possible, we calculated the probability that hospitals would have as many deaths as they did in 1986 and 1987, based on HCFA's published confidence intervals for death rates (for details, see Appendix A). We then calculated Pearson correlations across more than 5000 U.S. hospitals between the probabilities calculated using each of the targeting methods, for example, the correlation of our 30-day targeting probabilities for CHF using 1984 data with HCFA's targeting probabilities for chronic heart disease using 1986 data.

SAMPLING

For logistic reasons, we confined the sample to four states (California, Illinois, Minnesota, and New York), which together had 20 percent of U.S. hospitals and 22 percent of Medicare hospitalizations. Power calculations showed that a sample of 350 patients in each of the four targeted/untargeted dead/alive cells for each of the two conditions would be adequate. For a quality of care measure with standard deviation 1.0, we could expect to detect a 0.15 point difference in average quality between targeted and untargeted hospitals, or between dead and alive discharges, in 80 percent of repeated samples using a one-tailed test at the 0.05 significance level.

We drew a systematic random sample of discharges in the following manner. Our sample frame consisted of Medicare claims records arranged into eight lists. There was a separate list for each condition for each of its four cells: targeted or untargeted hospitals based on inpatient deaths, and dead or alive patients at time of discharge; for example, one cell included heart failure patients discharged dead from inpatient targeted hospitals. We sorted each list by state and hospital; within each hospital, we listed patients in random order and sampled systematically from that list.

We sampled based on inpatient deaths because that is what HCFA was using for its mortality data release at the time we drew the sample. HCFA subsequently shifted to analyzing 30-day deaths. Our sample proved useful for analyzing 30-day deaths as well. Analyzed using appropriate population weights, our sample yields unbiased estimates of differences between 30-day targeted and untargeted hospitals, albeit with slightly higher variance than the inpatient estimates. (A discussion and graphical comparison of inpatient compared with 30-day

death measures is in Appendix F. Additional information on sampling methods is in Appendix B.)

SIMULATION

To determine how much of the variation in death rates could have resulted from random variation or selection effects, we simulated hospital deaths in the four study states on the null hypothesis that the probability of death for each hospital was the age-sex-race standardized value d described under "targeting" above. Within each hospital, we simulated death for each of the n patients with probability d, and summed the simulated deaths, m^*. We then calculated the binomial probability of having as many as m^* deaths, $p^*(d,n,m^*)$ and ranked hospitals in order of p^*. We counted the h hospitals with the lowest values of p^* as targeted in the simulations, where h is the number of hospitals that were actually targeted—a number that differed for congestive heart failure and acute myocardial infarction and for inpatient or 30-day targeting.

We calculated death rates in the simulated targeted hospitals as a group (by summing m^* and n over those hospitals) and in simulated untargeted hospitals as a group. We repeated the process 100 times and averaged the simulated death rates over the 100 repetitions.

The simulations used administrative data, which turned out upon medical records abstraction to include a substantial number of cases that did not really belong in the study population for one reason or another as discussed below (Table 3). We adjusted both actual and simulated death rates in proportion to sample attrition so that both would represent the actual study population. (See Appendix C for additional information on simulation methods.)

MEDICAL RECORDS DATA

To collect detailed data on severity of illness and quality of care, we developed separate detailed abstraction forms for the two conditions (Kahn et al., 1988; Kosecoff et al., 1988), and contracted with local peer review organizations (PROs) in the four states to do the abstraction. We trained nurses and medical records abstractors in the use of the abstraction forms. The PROs asked the hospitals to send them complete photocopies of the sampled records. The records were abstracted at each PRO and the completed abstraction forms were sent to RAND, where selected items were reviewed first by a nonphysician to ensure completeness, legibility, and internal consistency of the

abstracted data. All of the abstracts were then reviewed by the physician principal investigator who determined for each case whether the principal diagnosis of congestive heart failure or acute myocardial infarction was accurately coded and verified other exclusionary criteria. We excluded patients if surgery occurred during the admission, if the patient had metastatic cancer or cancer under active treatment with radiation or chemotherapy, or if the patient had been transferred from another acute care hospital. We also excluded patients for whom the principal diagnosis of congestive heart failure or acute myocardial infarction was coded incorrectly.

We used the abstracted data to calculate disease-specific measures of severity of illness and quality of care. The measures are those developed by the RAND prospective payment system study (Kahn et al., 1990b).

The severity measure is a weighted sum of APACHE (Knaus et al., 1986) and other items including rescaled systolic blood pressure, the results of laboratory tests, and an inventory of chronic morbid and comorbid disease markers. It has been shown to predict 12 percent of the variance in deaths for patients with congestive heart failure and 22 percent for myocardial infarction (Keeler et al., 1990).

The quality score measures the process of care based on an explicit set of processes that should be done, including physician and nurse examination and history taking and the use of diagnostic, therapeutic, and intensive services. It is branched, i.e., different criteria apply to different patients. It is disease-specific and standardized to reflect differing levels in difficulty of complying with a criterion. Most important, it has been shown to be valid at the patient level, i.e., increased scores on this process scale result in lower probability of death (Kahn et al., 1990a).

ANALYSIS

Because we oversampled patients in targeted hospitals and patients discharged dead, we faced the question of whether (or when) to do population weighted analyses. We view the decision as an attempt to minimize mean squared error by striking a balance between potential bias and unnecessary variance in our estimates. If values (means, regression coefficients, or whatever) differ between sampling categories, then unweighted estimates will be biased estimates of true population values. Population weighted estimates will be unbiased, but they will have higher variance than the unweighted estimates. Thus when estimates do not differ much between sampling subgroups (and

consequently weighted estimates do not differ much from unweighted estimates), unweighted estimates are apt to be better because they have lower variance. On the other hand, when the estimates do differ importantly between sampling subgroups (and weighted and unweighted estimates differ), then weighted estimates will be better if one is interested in averages over subgroups, and separate estimates for the subgroups are more likely to be of interest. (See Appendix B for additional discussion of the relative merits of weighted and unweighted estimates.)

There is a further complication relevant to estimates of death rates or regressions with death as the dependent variable: Because we oversampled dead discharges, ours is a so-called "choice based" sample. (The name is unfortunate in this context, because death is not ordinarily a choice. The name arose in studies of travel demand in which people who chose to go by bus were oversampled relative to those who drove their own cars, and the data were used to estimate models of modal choice. The statistical issues are, however, the same in our case.) Manski and Lerman (1977) present sufficient conditions for consistent estimation using choice based samples. Briefly, they show that population weighted estimates are consistent. Also, they report a result of McFadden which states that unweighted logistic regression estimates are consistent (except for the estimated constant term). They do not mention, but it is obviously also true, that unweighted estimates are unbiased when the true values of the coefficients do not differ between sampling categories.

In general, we report unweighted regressions (ordinary least squares, logistic regression, and Cox proportional hazards estimates) here, but only after comparing them with weighted regressions and confirming that there are no substantial differences in the estimated coefficients (Appendix D). When we did find substantial differences between unweighted and weighted regression results, we took that to mean that the regression coefficients differed between sampling subpopulations, and we then reported separate regressions for the subpopulations.

In comparing mean values of quality of care, severity of illness, and other important variables in targeted compared with untargeted hospitals, we report population weighted estimates. There are good reasons to expect the levels of these variables to differ between dead and alive discharges, and the estimates of individual cell means confirm that such differences exist. Hence population weighted estimates are needed to avoid serious bias.

We used the Cox estimates to explore the effect of differences in severity of illness, quality of care, and length of stay on death rates. To do so we first adjusted the baseline hazard so that the weighted

average probability of death predicted using actual values for individual patients in the estimated Cox model equaled the observed death rate. Holding the baseline hazard constant, we then calculated "what if" death rates by changing individual patient values for quality or severity in targeted hospitals by constant amounts that undid the estimated differences between targeted and untargeted hospitals. For example, if we estimated that AMI patients in 30-day targeted hospitals had severity scores that averaged 3.0 higher than those in untargeted hospitals, we would use the Cox estimates to predict death rates on the counterfactual assumption that each sample patient in a targeted hospital had a severity score 3.0 points lower than actually observed. (See Appendix G for additional detail on the use of the Cox estimates to explore reasons for differences in death rates.)

RETARGETING

To understand whether our results are sensitive to the targeting method chosen (i.e., hospitals with $p < 0.05$ of having as many deaths as observed compared with all others), we changed how we classified hospitals and thereby tested other comparisons.

First, we looked for differences *within* our targeting categories. We subdivided targeted hospitals into those with $p < 0.01$ of having as many deaths as observed and those with $p \geq 0.01$, and subdivided untargeted hospitals into those with $p \geq 0.50$ and those with $p < 0.50$. We did this using both inpatient and 30-day targeting probabilities.

Second, we compared only the "best" and "worst" hospitals in our original sample. We defined the "best" hospitals as those that had lower than expected death rates (not necessarily significantly lower), thus excluding small hospitals with high death rates but too few patients to achieve statistical significance, and also excluding some larger hospitals with moderately high death rates. We defined the "worst" hospitals as those with $p < 0.01$ of having as many deaths as observed (which is also one of the categories in the first retargeting). We retargeted "best" and "worst" hospitals using both inpatient and 30-day death rates.

Third, we took advantage of HCFA's 1988 analysis of 1986 hospital data (Bowen and Roper, 1988). Our targeting method is similar to HCFA's method in that both use administrative data to target hospitals that are unlikely to have had as many deaths as they did if they were similar to other hospitals. But the methods differ in several respects. HCFA adjusts for severity; we adjusted only for age, sex, and race. (HCFA's 1989 method of adjusting for severity is a substantial

improvement on their earlier methods, but their 1989 data release was not available when this study was done.) HCFA's two conditions that correspond most closely to congestive heart failure and acute myocardial infarction (severe chronic and severe acute heart disease) are more broadly defined than our conditions. HCFA uses only the last discharge of the year, whereas we use all discharges; thus HCFA's death rates are higher than ours. HCFA's methods for setting cut points for outliers have changed over the years; for both 1986 and 1987, they yielded a smaller number of hospitals targeted for chronic heart disease than did our methods. Although we did not attempt to replicate HCFA's current targeting method on our 1984 data, we did—and this is our third alternative targeting method—reclassify our hospitals (1984 data) as targeted if they had $p < 0.05$ of having as many deaths as observed in HCFA's analysis of 30-day deaths during 1986.

Fourth, we developed an ad hoc three-year targeting method in an attempt to take advantage of random effects averaging out over time. Specifically, we multiplied together the probabilities that a hospital would have as many deaths as it did from our 1984 30-day death analysis, and HCFA's 1988 analyses of 1986 and 1987 data. We then ranked hospitals by the result of that computation, and counted as targeted the same number of hospitals from the top of the list that our 30-day method targeted in 1984.

Fifth, we attempted to see if adjusting for severity of illness as determined from clinical review of medical records might yield more precise targeting of hospitals. To shed light on this issue, we looked for differences in quality of care received by patients who lived when expected to die ("miracles") and those who died when expected to live ("disasters"). We said that a patient was expected to die if the probability of death predicted based on severity score alone was greater than 0.50; a patient was expected to live if that predicted probability was less than 0.50. Presumably, hospitals targeted using clinical severity-adjusted death rates would tend to have higher numbers of "disasters;" if we were to find that "disasters" received worse care, that would strengthen the case for adjusting the mortality data using a clinical severity measure.

III. RESULTS

HOSPITAL TARGETING WITH ADMINISTRATIVE DATA: NATIONWIDE CORRELATIONS

Table 1 shows Pearson correlations, across more than 5000 U.S. hospitals, among the probabilities that a hospital would have as many deaths as actually observed, calculated for each targeting method, condition, and year. The correlations are all positive. Certain correlations of particular interest are set off by braces {} or parentheses (). Additional results for two NOS targeting methods plus two others, and for a score of other conditions, are in Chassin et al. (1989). (NOS abbreviates Nonintrusive Outcomes Study, our name for the project we are reporting here. It is sometimes convenient to use the abbreviation to distinguish between NOS targeting and HCFA targeting.)

The correlations between NOS inpatient probabilities and 30-day probabilities (set off by parentheses) are quite high: 0.70 for CHF and 0.80 for AMI. The correlations between years for the same condition and the same targeting method (set off by braces) are 0.23 for chronic heart disease in 1986 and 1987, and 0.29 for acute heart disease in 1986 and 1987. Disregarding the differences in the way the conditions are defined, one can also compare NOS 30-day targeting of CHF in 1984 with HCFA targeting of chronic severe heart disease in 1986 and 1987, and similarly for NOS AMI and HCFA acute severe heart disease. The correlations here range from 0.15 to 0.19. One might expect them to be lower than the 1986/1987 correlations within disease both because of differences in disease definitions and because of the longer time interval between observations. (See Appendix A for additional details and results concerning outcome targeting nationwide, including relationships between targeting probabilities and various hospital characteristics.)

ACTUAL AND SIMULATED DEATH RATES IN FOUR STUDY STATES

In our four study states 1137 hospitals had at least one Medicare admission for a patient with congestive heart failure and 1121 hospitals had at least one admission for a patient with a myocardial infarction. Using our original targeting method (i.e., $p < 0.05$), 13 percent of hospitals were targeted using inpatient deaths and 7 percent using deaths

Table 1

NATIONWIDE CORRELATIONS AMONG PROBABILITIES OF HAVING AS MANY DEATHS AS ACTUALLY EXPERIENCED, FOR VARIOUS TARGETING METHODS
(N = 5348 hospitals)

	(1)	(2)	(3)	(4)
Chronic Heart Disease				
(1) NOS inpatient, CHF, 1984	1.00			
(2) NOS 30-day, CHF, 1984	(0.70)	1.00		
(3) HCFA 30-day, chronic, 1986	0.16	{0.19}	1.00	
(4) HCFA 30-day, chronic, 1987	0.13	{0.16}	{0.23}	1.00
Acute Heart Disease				
(1) NOS inpatient, AMI, 1984	1.00			
(2) NOS 30-day, AMI, 1984	(0.80)	1.00		
(3) HCFA 30-day, acute, 1986	0.14	{0.16}	1.00	
(4) HCFA 30-day, acute, 1987	0.13	{0.15}	{0.29}	1.00

SOURCE: Calculated from administrative data.

NOTES: Numbers in parentheses denote correlations across death measures (same year, same condition). Numbers in braces denote correlations across years (same death measure, same condition). Table A.2 presents the full 8x8 correlation matrix, including correlations across conditions. Table A.3 shows correlations among the logit transforms of the probabilities, which are almost the same as the correlations among the probabilities themselves.

within 30 days of admission for patients with congestive heart failure; 9 percent of hospitals were targeted using inpatient deaths and 6 percent using 30-day deaths for patients with an acute myocardial infarction. Of the hospitals targeted for one condition, using inpatient deaths, 22 percent were also targeted for the other condition. Using 30-day deaths, the overlap was about 17 percent.

Table 2 shows that death rates in targeted hospitals are substantially higher than those in untargeted hospitals, ranging from 40 percent higher for congestive heart failure 30-day deaths to almost 100 percent higher for congestive heart failure inpatient deaths. For acute myocardial infarction patients, targeted hospitals have about 50 percent higher actual death rates, regardless of whether deaths are counted in the hospital or within 30 days of admission. The simulation results show how much of the difference in death rates could result solely from the targeting method even if hospitals did not differ in quality of care or case mix (beyond age-sex-race). Even though differences in death rates in the targeted and untargeted hospitals are

Table 2

ACTUAL AND SIMULATED DEATH RATES FOR 1137 HOSPITALS TREATING CHF PATIENTS AND 1121 HOSPITALS TREATING AMI PATIENTS[a]

	CHF Patients		AMI Patients	
	Inpatient Deaths	30-Day Deaths	Inpatient Deaths	30-Day Deaths
Actual deaths per 100 patients in targeted hospitals	15.4	17.6	30.2	34.1
Actual deaths per 100 patients in untargeted hospitals	7.9	12.6	20.0	23.2
Targeted minus untargeted actual deaths per 100 patients	7.4	5.0	10.2	10.9
Targeted minus untargeted simulated deaths per 100 patients	5.1 (0.2)	4.1 (0.3)	6.3 (0.5)	6.1 (0.6)
% of actual difference in death rates between targeted and untargeted hospitals that could be due to random variation	69	82	62	56

[a]Death rates are for hospitals in four states, October 1983 through September 1984. Simulated values are means from 100 trials, with standard deviations in parentheses.

statistically significant, random variation and the selection of targeted hospitals could account for a large share, between 56 and 82 percent, of the differences. The remaining nonrandom components of the death rate differences between targeted and untargeted hospitals are both clinically important and highly significant statistically (Appendix C). For example, at 30 days postadmission an additional 4.8 deaths per 100 patients admitted with myocardial infarction are unexplained after allowing for the way targeted hospitals were selected; the corresponding figure for congestive heart failure patients is 0.9 deaths per 100 patients admitted. (For additional information on the simulation method and results, see Appendix C.)

VALIDATING SEVERITY OF ILLNESS AND QUALITY OF CARE MEASURES

We used a sample of medical records to determine the extent to which the differences in targeted and untargeted death rates resulted from differences in severity and quality. Table 3 summarizes sample

Table 3

POPULATION AND SAMPLE COUNTS BY SAMPLING CATEGORY AFTER SAMPLE ATTRITION

	CHF Patients		AMI Patients	
	Untargeted Hospitals	Targeted Hospitals	Untargeted Hospitals	Targeted Hospitals
	Hospitals			
Four states	992	145	1017	104
Sampled	533	141	525	104
% participating	516 (97)	137 (97)	511 (97)	100 (96)
	Patients			
Four states	65,702	16,465	74,844	7,322
Sampled	800	800	800	800
Less:				
Hospital not identifiable[a]	0	14	0	18
Hospital refused	23	12	15	59
Coding errors[a]	124	124	116	115
Claims data errors and other exclusions[a]	40	52	26	37
Failure to obtain usable copy of sampled record	30	55	21	43
Usable data (% of true records[b])	583 (92)	543 (89)	622 (95)	528 (84)

[a]Unavoidable attrition.
[b]After excluding unavoidable attrition.

attrition. Of the 3200 sampled patients, 32 were from hospitals that we could not identify from HCFA data. We obtained 97 percent participation by sampled hospitals. This resulted in 2 percent of congestive heart failure and 5 percent of acute myocardial infarction patients being excluded from our sample. Upon examining medical records, we found that 248 (16 percent) of sampled congestive heart failure and 231 (14 percent) of acute myocardial infarction patients had to be excluded because of coding errors; that is, the intended condition was not the true principal diagnosis. Coding errors tend to decrease the precision of targeting based on administrative data; they may be less prevalent now than in 1984, the first year of prospective payment for DRGs.

In addition, 92 congestive heart failure patients and 63 acute myocardial infarction patients were excluded for other reasons, such as claims data errors or the patients died in the emergency room. Finally, 85 congestive heart failure patients and 64 acute myocardial infarction

patients were excluded because the hospital was unable to locate the sampled admission. We thus obtained complete data on 1126 patients (90 percent) of the 1246 patients with congestive heart failure who were eligible for the study after excluding those ineligible because of claims or coding errors and on 1150 patients (89 percent of eligibles) with myocardial infarction. (See Appendix B for more information on sample attrition.)

Table 4 summarizes our patient level results on the effect of severity of illness and quality of care on probability of death, together with some other important relationships involving age, do not resuscitate (DNR) orders, and length of hospital stay. The results help to establish the validity of the severity and quality measures. The severity and quality measures, as mentioned in Sec. II, were developed in a national study that examined the impact of diagnosis related groups (Kahn et al., 1990b).

Column (1) of the table shows logistic regressions for the presence of a DNR order dated on the admission day, regressed on patient age and severity of illness. As expected, sicker patients are more likely to have DNR orders written. The severity measure includes age as one factor insofar as it affects the probability of death. It is thus perhaps a little surprising that age independently increases the probability of a DNR order, after controlling for severity of illness.

Column (2) shows ordinary least squares regressions of quality scores on age, severity of illness, and a DNR indicator variable. Higher quality scores correspond to better care, so the regressions show that older patients, sicker patients, and patients with a DNR order all tend to get worse care. Even though older, sicker, and DNR are all correlated, they have independently significant effects on quality.

Columns (3) and (4) show ordinary least squares regressions of the natural logarithm of length of hospital stay (log(LOS)) on severity, DNR, and quality, separately for patients discharged alive (column (3)) and those discharged dead (column (4)). For CHF patients discharged dead, higher severity and DNR orders both decrease the length of stay (and time to death), whereas better quality of care increases length of stay (and time to death). The relationship for AMI patients discharged dead is similar, except that the effect of DNR at admission is not significant. For patients discharged alive (both CHF and AMI), greater severity of illness is associated with longer hospital stays, and the effects of DNR and quality are relatively weak.

It makes sense that severity has a different effect on length of stay for patients who live and those who die. Those who die die faster the sicker they are. Those who live require longer hospitalization the sicker they are.

Table 4

REGRESSION RESULTS

	Logit DNR	OLS Quality	OLS log(LOS) Alive	OLS log(LOS) Deadin	Logit Deadin	Logit Dead30	Cox Deadin	Cox Dead30	Cox Dead30
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
CHF Patients									
Age	0.36	−0.12	—	—	—	—	—	—	—
	(2.3)	(−3.8)							
Severity	8.43	−1.06	1.11	−2.14	13.43	12.96	5.47	7.22	5.83
	(6.3)	(−3.3)	(3.1)	(−4.2)	(13.6)	(13.3)	(12.1)	(15.9)	(12.7)
DNR	—	−0.55	0.05	−0.64	1.32	1.85	0.87	0.95	0.94
		(−4.5)	(0.2)	(−4.3)	(3.2)	(4.1)	(5.9)	(6.4)	(6.4)
Quality	—	—	−0.01	0.14	−0.29	−0.24	−0.12	−0.13	−0.10
			(−0.5)	(3.2)	(−3.8)	(−3.1)	(−2.7)	(−2.8)	(−2.2)
Home	—	—	—	—	—	—	—	—	−2.62
									(−11.7)
Constant	−8.94	1.32	1.84	3.03	−5.00	−4.98	—	—	—
	(−7.0)	(5.4)	(15.7)	(14.5)	(−14.1)	(−14.1)			
R-square	0.05	0.06	0.02	0.11	0.27	0.27	—	—	—
Observations	1126	1126	608	518	1126	1126	1126	1126	1714
AMI Patients									
Age	0.79	−0.18	—	—	—	—	—	—	—
	(3.9)	(−5.3)							
Severity	4.88	−0.83	0.48	−2.10	7.77	7.53	4.58	4.74	4.59
	(6.1)	(−5.5)	(2.6)	(−9.9)	(14.0)	(13.8)	(18.3)	(19.1)	(18.5)
DNR	—	−0.92	−0.13	−0.09	1.74	1.45	0.34	0.44	0.42
		(−7.1)	(−0.4)	(−0.6)	(2.7)	(2.6)	(2.0)	(2.6)	(2.4)
Quality	—	—	0.05	0.17	−0.06	−0.02	−0.16	−0.11	−0.13
			(2.0)	(4.1)	(−0.7)	(−0.2)	(−3.4)	(−2.4)	(−2.9)
Home	—	—	—	—	—	—	—	—	−2.26
									(−7.4)
Constant	−11.39	1.88	2.42	2.08	−2.32	−2.25	—	—	—
	(−6.8)	(7.3)	(53.3)	(23.1)	(−13.6)	(−13.4)			
R-square	0.06	0.13	0.02	0.22	0.27	0.25	—	—	—
Observations	1149	1149	596	553	1149	1149	1149	1149	1727

NOTE: Column headings show estimation method (top line), dependent variable (middle line), and subpopulation if applicable (third line); t-statistics are in parentheses. The higher number of observations for the Cox estimates that include the home indicator variable (column (9)) is an artifact of the estimation method, which requires replicating observations for patients discharged alive less than 30 days after admission to create one observation before and one after discharge. Severity and DNR are scaled differently here than in Tables 5 and 6 (see Table D.1).

Columns (5) and (6) show logistic regressions of inpatient (Deadin) and 30-day death (Dead30) on severity, DNR, and quality. For CHF, the relationships are all highly significant in the expected directions. For AMI, greater severity and DNR orders are both significantly associated with a higher probability of dying, but the effect of quality of care is not statistically significant.

The logistic regression estimates for inpatient death are suspect because variable length of stay is not accounted for. Holding everything else constant, hospitals that keep their patients longer before discharging them will have more of them die in the hospital. Yet it is not legitimate simply to include length of stay as an explanatory variable in these regressions, because length of stay not only affects, but is affected by, inpatient death: Death truncates length of stay.

The Cox proportional hazards estimates in columns (7), (8), and (9) are probably a better way to estimate the death equations. This is because they estimate the relative probability of dying on any given day after hospitalization for the cohort of patients that are still alive on that day, rather than the probability of dying during the highly variable period of hospitalization. The Cox estimates have the additional advantage of using information on the length of time that a patient survives, not just on whether he eventually dies or not, and so should produce more efficient estimates.

The Cox estimates for inpatient deaths in column (7) keep each patient in the cohort being observed until hospital discharge. The 30-day death estimates in columns (8) and (9) truncate the observation of each patient at death or at 30 days following admission, whichever comes first. The 30-day estimates are done two ways, with and without a new independent variable, "Home," which equals one on days after a patient has been discharged from the hospital, and zero while he is still in the hospital. Its highly significant negative coefficient in column (9) indicates that patients are much less likely to die on any given day following admission to the hospital if they have gotten well enough to be discharged.

The other coefficients are all significant and have the expected signs (i.e., validity of the severity and quality of care measures is established). Greater severity of illness and DNR status independently increase the risk of death; better quality of care lowers it. For example, for the 30-day congestive heart failure model, patients at the 25th percentile of severity and at the median for DNR and quality of care have a predicted 30-day death rate of 9.0 per 100 admissions; those at the 75th percentile of severity have 17.4 predicted deaths. Correspondingly, patients at the 25th percentile of quality have 13.4 predicted deaths, and those at the 75th percentile of quality have 11.9 predicted deaths.

For CHF patients, the Cox estimates do not differ substantially from the logistic regression estimates. (Only the relative magnitudes of the coefficients matter, and these do not differ much between the two estimation methods.) For AMI patients, the Cox estimates are also similar to the logistic regression estimates, except that good quality significantly reduces the risk of death in the Cox equation but has no significant effect in the logistic equation.

Alternative estimates for all of the equations in Table 4 (shown in Appendix D) indicate that the results in Table 4 are not artifacts of the particular model specifications reported here. The results shown do not change substantially when additional variables are added to the regressions to control for geographic site or hospital characteristics. Nor do they change substantially when population weighted estimates are substituted for these unweighted estimates. (Weighted and unweighted estimates do differ when the length of stay equations are estimated for the entire population. Those differences indicate that these relationships differ among sampling categories, and that is what led us to estimate separate equations for patients discharged alive and those discharged dead.) Nor do they differ when the estimates are done separately for patients who died compared with those who lived, or for targeted compared with untargeted hospitals, except as just noted.

COMPARING TARGETED AND UNTARGETED HOSPITALS

Tables 5 and 6 compare targeted and untargeted hospitals in terms of average severity of illness, quality of care, percentage of patients who had DNR orders written at admission, and length of stay. The comparisons are presented for both inpatient and 30-day targeting. There are separate comparisons for dead and alive patients (at discharge for inpatient targeting, at 30 days postadmission for 30-day targeting), as well as, after reweighting for the sampling strategy, averages over all patients.

Significant differences between targeted and untargeted hospitals are marked with plus signs (for differences that go in the expected direction, i.e., targeted hospitals have lower quality of care or more severely ill patients) or minus signs (for differences that go in the unexpected direction). There are only spotty differences, and they go in the unexpected direction as often as not. Congestive heart failure patients who died within 30 days of admission received significantly worse care in 30-day targeted hospitals than in untargeted hospitals, but in all the

Table 5

DIFFERENCES IN SEVERITY OF ILLNESS, DNR STATUS, QUALITY OF CARE, AND LENGTH OF STAY BETWEEN TARGETED AND UNTARGETED HOSPITALS FOR CHF PATIENTS

	Inpatient Targeting		30-Day Targeting	
	Untargeted Hospitals	Targeted Hospitals	Untargeted Hospitals	Targeted Hospitals
Patients in sample				
Alive	318	290	526	109
Dead	265	253	402	89
Total	583	543	928	198
Severity score				
Alive	32.00	30.91	31.51	31.22
	(0.44)[a]	(0.37)	(0.32)	(0.61)
Dead	41.37	39.75 –	40.39	40.37
	(0.56)	(0.56)	(0.45)	(0.88)
Weighted average[b]	32.75	32.27	32.63	32.82
	(0.34)	(0.32)	(0.27)	(0.54)
DNR status at admission (%)				
Alive	2.20	0.34 –	1.02	0.00 –
	(0.82)	(0.34)	(0.44)	(0.00)
Dead	16.60	4.74 – –	13.34	23.22 +
	(2.29)	(1.34)	(1.70)	(4.50)
Weighted average[b]	3.35	1.02 – –	2.57	4.08
	(0.75)	(0.43)	(0.52)	(1.41)
Quality of process score[c]				
Alive	0.10	0.22	0.12	0.23
	(0.05)	(0.05)	(0.04)	(0.08)
Dead	–0.27	–0.14	–0.13	–0.42 +
	(0.06)	(0.06)	(0.05)	(0.13)
Weighted average[b]	0.07	0.16	0.09	0.11
	(0.04)	(0.04)	(0.03)	(0.07)
Length of stay (days)				
Alive	9.72	13.24 ++	10.60	13.28
	(0.39)	(1.29)	(0.39)	(3.04)
Dead	9.20	18.78 ++	8.84	8.41
	(0.58)	(1.42)	(0.32)	(0.71)
Weighted average[b]	9.68	14.10 ++	10.38	12.42
	(0.29)	(0.95)	(0.28)	(2.06)

[a]Standard errors are in parentheses. Significant differences between untargeted and targeted hospitals are marked as follows: ++ $p < 0.01$, expected direction; + $p < 0.05$, expected direction; – $p < 0.05$, unexpected direction; and – – $p < 0.01$, unexpected direction.

[b]Number was reweighted to reflect the fact that dead patients were oversampled relative to live patients.

[c]Higher score is better care; see the text and Kahn et al. (1990a).

Table 6

DIFFERENCES IN SEVERITY OF ILLNESS, DNR STATUS, QUALITY OF CARE, AND LENGTH OF STAY BETWEEN TARGETED AND UNTARGETED HOSPITALS FOR AMI PATIENTS

	Inpatient Targeting		30-Day Targeting	
	Untargeted Hospital	Targeted Hospital	Untargeted Hospital	Targeted Hospital
Patients in sample				
Alive	311	285	464	129
Dead	311	243	429	128
Total	622	528	893	257
Severity score				
Alive	21.63	20.92	21.61	22.68
	(0.65)[a]	(0.65)	(0.53)	(0.96)
Dead	38.09	39.15	35.84	38.62
	(1.01)	(1.20)	(0.87)	(1.65)
Weighted average[b]	24.92	26.44	24.91	28.11 ++
	(0.58)	(0.70)	(0.49)	(0.99)
DNR status at admission (%)				
Alive	0.96	0.00	0.95	0.00 –
	(0.56)	(0.00)	(0.45)	(0.00)
Dead	8.36	7.00	6.85	9.56
	(1.57)	(1.64)	(1.22)	(2.61)
Weighted average[b]	2.45	2.12	2.32	3.26
	(0.62)	(0.63)	(0.50)	(1.11)
Quality of process score[c]				
Alive	0.31	0.35	0.30	0.41
	(0.04)	(0.05)	(0.03)	(0.07)
Dead	0.05	0.05	0.12	0.15
	(0.06)	(0.06)	(0.05)	(0.08)
Weighted average[b]	0.26	0.26	0.26	0.32
	(0.03)	(0.04)	(0.03)	(0.05)
Length of stay (days)				
Alive	13.17	15.78 ++	13.64	15.20
	(0.35)	(0.65)	(0.33)	(1.09)
Dead	5.82	6.72	5.66	5.43
	(0.42)	(0.80)	(0.26)	(0.52)
Weighted average[b]	11.70	13.05 +	11.80	11.87
	(0.28)	(0.53)	(0.25)	(0.72)

NOTES: Standard errors are in parentheses. Significant differences between untargeted and targeted hospitals are marked as follows: ++ $p < 0.01$, expected direction; + $p < 0.05$, expected direction; – $p < 0.05$, unexpected direction; and — $p < 0.01$, unexpected direction.

[a]Standard errors are in parentheses.

[b]Number was reweighted to reflect the fact that dead patients were oversampled relative to live patients.

[c]Higher score is better care; see the text and Kahn et al. (1990a).

other quality comparisons for both congestive heart failure and acute myocardial infarction patients, targeted hospitals were as good as or better than untargeted hospitals, although never significantly so. Acute myocardial infarction patients in 30-day targeted hospitals were significantly sicker overall than those in untargeted hospitals, but that is the only significant severity comparison; for congestive heart failure patients the nonsignificant trends are in the unexpected direction.

Average quality scores and, especially, average severity scores appear to be estimated quite precisely in Tables 5 and 6, and the estimated differences between targeted and untargeted hospitals appear to be quite small. Still, it is worthwhile to investigate explicitly the importance of the estimated differences, and the importance of the uncertainty in the estimated differences, in terms of their implied effects on death rates. Table 7 summarizes such an investigation. The first step was to calculate 95 percent confidence intervals for differences in severity, quality, DNR, and length of stay in targeted minus untargeted hospitals, based on information in Tables 5 and 6 (the confidence intervals are in Tables G.1 and G.2). Then we used the Cox estimates to predict several "what if" death rates, based on those differences. The effects in Table 7 are based on those predicted "what if" deaths rates (as described fully in Appendix G).

The effects of estimated differences in quality are small, and what differences there are tend to favor targeted hospitals. That is, targeted hospitals from Tables 5 and 6 have *better* estimated average quality (except for acute myocardial infarction patients in inpatient targeted hospitals), so undoing the difference would *increase* their death rates. Moreover, even at the lower bound of the confidence intervals for inpatients with a myocardial infarction, where quality is worse in targeted than in untargeted hospitals, poorer quality would contribute, if that result were true, just 0.29 deaths per 100 admissions to excess deaths in targeted hospitals.

The estimated differences in severity also have fairly small effects on death rates for congestive heart failure patients. For myocardial infarction patients, however, higher average severity in targeted hospitals has a substantial effect on death rates. For example, Table 6 shows that acute myocardial infarction patients in 30-day targeted hospitals averaged 3.2 point higher severity scores than those in untargeted hospitals. If all such patients had 3.2 lower severity scores, there would be no difference in average severity between targeted and untargeted hospitals and the death rate in targeted hospitals predicted by the Cox estimates would decrease by 2.8 deaths per 100 admissions. At the upper bound of the confidence interval, the patients in targeted hospitals had 5.4 point higher severity scores. If all such patients had

Table 7

DIFFERENCES IN DEATH RATES BETWEEN TARGETED AND UNTARGETED HOSPITALS THAT CORRESPOND TO ESTIMATED DIFFERENCES IN SEVERITY, DNR, QUALITY, AND LENGTH OF STAY
(Deaths per 100 admissions)

Explanatory Variable	Inpatient Deaths[a]	30-Day Deaths[a]
	CHF Patients	
Due to severity difference	−0.25	0.16
95% confidence interval	(−0.72, 0.24)	(−0.79, 1.18)
Due to DNR difference	−0.19	0.16
95% confidence interval	(−0.33, −0.05)	(−0.16, 0.49)
Due to quality difference	−0.11	−0.04
95% confidence interval	(0.01, −0.23)	(0.19, −0.26)
Due to length of stay difference	1.87	—
95% confidence interval	(0.49, 3.21)	
	AMI Patients	
Due to severity difference	1.21	2.83
95% confidence interval	(−0.21, 2.70)	(0.89, 4.87)
Due to DNR difference	−0.02	0.08
95% confidence interval	(−0.13, 0.09)	(−0.12, 0.29)
Due to quality difference	0.02	−0.12
95% confidence interval	(0.28, −0.25)	(0.12, −0.36)
Due to length of stay difference	1.00	—
95% confidence interval	(−0.09, 1.81)	

[a]A negative sign means that if actual differences in severity, DNR status at admission, or quality were eliminated between targeted (high mortality) hospitals and untargeted hospitals, death rates at targeted hospitals would increase.

severity scores 5.4 lower than observed, that would have lowered targeted hospital death rates by 4.8 percentage points.

Another possibility is that the excess death rate in inpatient targeted hospitals results in part from longer average patient stays in those hospitals. Congestive heart failure patients stay on average more that four days longer in inpatient targeted hospitals (14.1 days compared with 9.7), and myocardial infarction patients stay more than one day longer (13.1 compared with 11.7). If CHF patients stayed in targeted hospitals only 9.7/14.1 = 0.69 times as long, so that targeted and untargeted hospitals had the same average length of stay (and the underlying risk of dying on each day of the hospitalization stayed the same), targeted hospitals would have had 1.9 fewer deaths per 100

admissions. Corresponding adjustments that undo the length of stay differences at the 95 percent confidence bounds on those differences yield a calculated reduction in deaths in targeted hospitals of 0.5 to 3.2 per 100 admissions. For AMI inpatient deaths, the effect of length of stay is smaller but still substantial: 1.0 deaths per 100 admissions, with a range of −0.1 to 1.8.

EXPLAINING DIFFERENCES IN DEATH RATES BETWEEN TARGETED AND UNTARGETED HOSPITALS

Table 8 is a heuristic "explanation" of higher death rates in targeted hospitals. It pulls together some already-discussed numbers that show how much of the spread between targeted and untargeted death rates could result from randomness or the way targeted hospitals were selected, and how much from measured differences in severity, quality, DNR, and length of stay. Consider first the 30-day targeting method for acute myocardial infarction patients. From Table 1, we know that targeted hospitals have a 30-day death rate that is 10.9 percentage points higher than untargeted hospitals, and that 6.1 percentage points of that difference could result from random variation and the way hospitals were targeted, even if all hospitals provided the same quality of care to patients that were identical except for age, sex, and race. From the calculations in Table 7, we know that the estimated differences in average severity, DNR status at admission, and quality could account for an additional change of 2.8, 0.1, and −0.1 percentage points in the death rate in targeted hospitals. That leaves a 2.0 percentage point gap that can not be accounted for either by the way targeted hospitals were selected or by measured differences between targeted and untargeted hospitals.

There is sufficient uncertainty in our estimate of the difference in average severity between targeted and untargeted hospitals that, if that difference were at the upper limit of the 95 percent confidence interval, severity could account for an additional 2.0 percentage points of difference in death rates (4.87 − 2.83 = 2.04 from Table 7). Thus if actual severity differences were at their 95 percent upper confidence bound, they would be sufficient to close the unexplained gap.

The situation is similar for inpatient targeting of hospitals treating either CHF or AMI patients. Estimated differences in severity of illness and length of stay account for a little over half of the gap between targeted and untargeted deaths rates that could not result from random binomial variation. Greater severity and length of stay differences that are still within the estimated confidence bounds for those differences

Table 8

EXPLAINING EXCESS DEATH RATES IN TARGETED COMPARED WITH UNTARGETED HOSPITALS
(Deaths per 100 admissions)

	CHF Patients		AMI Patients	
	Inpatient Targeting	30-Day Targeting	Inpatient Targeting	30-Day Targeting
Observed difference	7.4	5.0	10.2	10.9
Less:				
Expected due to binomial variation or selection effect	5.1	4.1	6.3	6.1
Expected due to measured differences in severity of illness	−0.3(ns)	0.2(ns)	1.2(ns)	2.8
Expected due to differences in DNR at admission	−0.2(ns)	0.2(ns)	−0.0(ns)	0.1(ns)
Expected due to measured differences in quality of care	−0.1(ns)	−0.0(ns)	0.0(ns)	−0.1(ns)
Expected due to differences in length of stay	1.9	—	1.0(ns)	—
Unexplained after binomial variation and measured differences	1.0	0.6	1.7	2.0

NOTE: (ns) indicates no significant difference between targeted and untargeted hospitals ($p > 0.05$). Minus sign indicates that lower severity, lower DNR, or higher quality in targeted hospitals contributes to the difference to be explained, rather than helping to explain the difference.

would be sufficient to close the gap. For CHF 30-day deaths, a small gap of only 0.6 deaths per 100 admission remains unexplained after allowing for selection effects and estimated systematic differences in severity, DNR, and quality.

The magnitude of the spread resulting from random selection effects depends, of course, on the targeting method used. We do not believe that it would be much changed for other methods that use administrative data only, but our results do suggest that for AMI, a 30-day targeting method that used clinical severity adjustment would reduce the spread resulting from selection effects by about 2.8 deaths per 100 patients. Better severity measures that may be developed in the future could reduce the spread even more.

Some people would argue that there is no place for randomness in medicine, that if we could measure disease severity perfectly at the cellular level, and target after adjusting for a perfect severity measure,

any remaining differences between death rates in targeted and untargeted hospitals would be due to the care that patients received. We agree, but we doubt that anything approaching a perfect severity measure will ever be available. If not, then it makes little or no practical difference whether the spread resulting from selection effects is truly random or simply unexplainable by anything we can measure.

RETARGETING

Table 9 compares average quality of care received by patients in alternatively targeted compared with untargeted hospitals. (Appendix E contains more detailed tables comparing these groups of hospitals on other dimensions, including severity of illness, fraction of DNR, and length of stay, and separately for dead and alive patients as well as the overall averages.) The results (in both Table 9 and Appendix E) are mixed, with the multiple year targeting method looking the most promising.

Comparing only the "best" (defined either as $p > 0.50$ or as lower than expected deaths) and "worst" ($p < 0.01$ of so many deaths) hospitals shows higher quality in targeted hospitals as often as not. Hospitals targeted at the 0.05 level in HCFA's 1988 analysis of 1986 data for chronic severe heart disease had significantly lower average quality for CHF patients in our 1984 sample, but the estimate is suspect because it is based on only 26 patients; the difference for AMI patients was not significant. Three-year targeting identifies significantly lower explicit quality in targeted hospitals treating CHF patients and a nearly significant trend in the same direction for AMI patients.

A considerable number of patients in our sample died even though their severity score was low, or lived despite a high severity score—more than 100 in each group for each condition. One might expect that the latter got better care than the former. That expectation is not borne out in our data. In three out of four comparisons, those who died unexpectedly had higher average quality scores than did those who lived unexpectedly, and for AMI 30-day deaths, they were significantly higher.

Table 9

DIFFERENCES IN QUALITY OF CARE BETWEEN TARGETED AND UNTARGETED HOSPITALS USING ALTERNATIVE TARGETING METHODS

	CHF Patients		AMI Patients	
	Untargeted Hospitals	Targeted Hospitals	Untargeted Hospitals	Targeted Hospitals
Best ($p > 0.50$) Compared with Worst ($p < 0.01$) Inpatient Targeting				
Patients in sample	313	370	345	226
Quality of process score[a]	0.15	0.21	0.23	0.36
Standard error	(0.05)	(0.05)	(0.05)	(0.05)
Best ($p > 0.50$) Compared with Worst ($p < 0.01$) 30-Day Targeting				
Patients in sample	433	96	419	95
Quality of process score[a]	0.11	−0.05	0.25	0.15
Standard error	(0.04)	(0.10)	(0.04)	(0.10)
Best (Lower Than Expected Deaths) Compared with Worst ($p < 0.01$) Inpatient Targeting				
Patients in sample	265	370	328	226
Quality of process score[a]	0.17	0.21	0.25	0.36
Standard error	(0.05)	(0.05)	(0.05)	(0.05)
Best (Lower Than Expected Deaths) Compared with Worst ($p < 0.01$) 30-Day Targeting				
Patients in sample	399	96	402	95
Quality of process score[a]	0.12	−0.05	0.25	0.15
Standard error	(0.04)	(0.10)	(0.04)	(0.10)
Hospitals Targeted by HCFA for 1986 ($p < 0.05$)				
Patients in sample	1080	26	984	141
Quality of process score[a]	0.11	−0.42 ++	0.28	0.18
Standard error	(0.03)	(0.16)	(0.03)	(0.07)
Three-Year Targeting				
Patients in sample	950	156	1006	117
Quality of process score[a]	0.13	−0.20 ++	0.28	0.13
Standard error	(0.03)	(0.08)	(0.03)	(0.08)
Patients Who Lived When Predicted to Die Compared with Those Who Died When Predicted to Live; Inpatient Deaths				
	"Miracles"	"Disasters"	"Miracles"	"Disasters"
Patients in sample	136	186	125	199
Quality of process score[a]	−0.08	−0.14	0.23	0.40
Standard error	(0.08)	(0.08)	(0.08)	(0.06)
Patients Who Lived When Predicted to Die Compared with Those Who Died When Predicted to Live; 30-Day Deaths				
Patients in sample	123	198	128	203
Quality of process score[a]	−0.08	0.10	0.23	0.44 −
Standard error	(0.08)	(0.07)	(0.08)	(0.06)

NOTE: Significant differences between untargeted and targeted hospitals are marked as follows: ++ $p < 0.01$, expected direction; − $p < 0.05$, unexpected direction.

[a]Higher score is better care; see the text and Kahn et al. (1990a).

IV. DISCUSSION

We stated our main study objectives were to determine (1) if hospitals with high death rates provide lower quality care or have more severely ill patients than do hospitals with lower death rates, and (2) how the probability of death at the patient level is related to severity of illness and quality of care.

With respect to the first objective, we determined that hospitals targeted with unexpectedly high age-sex-race-disease-specific death rates do not provide lower quality of care than do untargeted hospitals, and that any differences in quality of care that lie within estimated confidence bounds have minimal effects on death rates. We found that higher average severity for myocardial infarction patients in targeted hospitals accounts for almost 25 percent of the difference in 30-day death rates, but differences in severity of illness do not explain higher death rates for congestive heart failure patients.

With respect to the second objective, we determined that, at an individual patient level, higher severity of illness markedly increases the probability of death, and, to a lesser extent, better quality of care reduces the probability of death.

Finally, we are left with an unexplained excess of 1.0 or 0.6 deaths per 100 patients admitted with congestive heart failure for hospitals targeted on inpatient or 30-day death rates. For acute myocardial infarction patients, the figures are 1.7 and 2.0 excess deaths per 100 admitted patients. The excess for 30-day targeting could possibly result from misestimated severity and quality differences between targeted and untargeted hospitals, but the uncertainty in these estimates is not large enough to explain the excess for inpatient targeting. The excess for inpatient targeting could be the effect of unmeasured severity differences, unmeasured quality differences, or longer average lengths of stay in targeted hospitals.

Unmeasured severity of illness could be responsible for some of this excess, but another study has demonstrated that it will be difficult to improve the measurement of severity using just data in a medical record for patients with these two conditions (Keeler et al., 1990). If the hypothesis that hospital death rate differences are due to unmeasured severity is to be tested, then prospectively collected data from the patient or physician must be obtained.

Even the sophisticated explicit measures of quality used here certainly cannot capture all of the potentially important differences in hospital care, if for no other reason than that they measure the process of care predominantly during the first three days following admission. A previous study demonstrated a relationship between quality of hospital care and the hospital death rates, but only when using implicit physician judgment to measure quality (not preset criteria) (Dubois et al., 1987). That study found no significant relationship between death rates and a quality score explicitly calculated from medical record data based on preset criteria. Any defects in quality in targeted hospitals appeared to be in areas not easily assessed by explicit measurement.

If one believes that quality differs among hospitals and that it is important to detect the differences, the more important question is whether a targeting mechanism can be devised that better identifies hospitals providing lower quality care. For that reason, we retargeted the hospitals in our data set to take account of several possible limitations in our targeting method. One possibility is that a $p < 0.05$ cutoff for targeting (where p is the probability that the hospital would have as many deaths as it did) is not strict enough, or that differences are obscured by contrasting hospitals below the probability cutoff with *all* other hospitals (which may include small hospitals with high death rates and low quality care, but too few patients to reach the statistical significance necessary for targeting). Another possibility is that our method does not adequately control for severity of illness. HCFA's method is similar to ours but adjusts for severity as thoroughly as possible using administrative data; we adjusted only for age, sex, race, and disease. Even better severity adjustment could be attained by using clinical data. A third possibility is that random variation obscures any real differences in a single year analysis. When we retargeted to minimize or avoid these possible problems, the results were mixed, and except possibly for three-year targeting, did not add much to our original analysis.

In summary, our analyses of a representative sample of heart failure and myocardial infarction patients in four populous states have not produced much evidence that hospitals with higher than expected death rates based only on administrative data actually, on review of their medical records, deliver lower quality care. What can be done to improve our ability to identify such hospitals? There is some evidence that targeting hospitals with consistently high death rates over periods longer than one year may identify potential quality problems. Further research needs to be performed to identify the optimal targeting interval (two, three, or four years), but the time interval cannot be so long as to make the results irrelevant to current patient care. HCFA's

current mortality release reports results for three years of data, but they are analyzed one year at a time (Sullivan and Hays, 1989).

Our targeting method did not use all of the information available in the administrative data to control for severity of illness, only age, sex, race, and a clinical meaningful grouping of ICD-9 principal diagnoses. Thus it is closer to the method used by HCFA in its initial 1986 data release (Brinkley, 1986) (when this study was designed) than it is to HCFA's current method (Sullivan and Hays, 1989). Would targeting that included other severity adjustment measures that are available from administrative data have substantially affected the results of this study? We think not. Green et al. found that the severity adjustment used in HCFA's 1988 release (Bowen and Roper, 1988) explained only 2.5 percent of the variance in outcome on average for the five broadly defined diseases investigated, including 3.9 percent for severe acute heart disease (Green et al., 1990). HCFA's 1989 analysis, which controls for principal diagnosis grouped into homogeneous death rate categories (Sullivan and Hays, 1989), might explain on the order of 8 percent of the variance in our already fairly homogeneously defined acute myocardial infarction, and probably only 2 percent for our congestive heart failure. Thus even if we had used the most recent HCFA method to adjust for severity, we would have reduced the width of the targeting confidence intervals by no more than 4 percent or so, and the targeting probabilities calculated with and without the additional adjustment would have been highly correlated.

HCFA has substantially improved its targeting method over the years, and continues to improve it. But given our results, we believe that the improved methods should be tested to see if they are indeed targeting lower quality hospitals. First, such methods should be tested against simulation models to confirm that they are not just picking hospitals whose high death rates could result solely from random variation, as was the case for 26 of 48 conditions studied in Chassin et al. (1989). Second, the quality of care in targeted compared with untargeted hospitals should be compared using clinical data from medical records. Third, that comparison should include both implicit and explicit assessment of quality. Fourth, sufficient public discussion about both the targeting methods and results should occur so that perhaps their acceptability within the medical profession and the hospital community will increase (Berwick and Wald, 1990). Fifth, if targeting based on administrative data cannot be improved, then serious attention needs to be given to whether detailed data on severity of illness at time of admission should be collected routinely and nationally. If hospitals were targeted based on detailed severity data, would the targeting be more accurate? Would the additional accuracy be worth

the cost? And finally, if, even after clinical severity adjustment, mortality data do not make a very good screen for identifying low quality hospitals, direct collection of data on the quality of the process of care received by a sample of patients should be considered.

Appendix A

OUTCOME TARGETING NATIONWIDE

In this appendix, we describe and compare results of several different ways of targeting high death rate hospitals using administrative data. These methods include:

1. NOS inpatient targeting, applied to all admissions, during fiscal year 1984, of Medicare patients 65 years old or older, to U.S. acute care hospitals, for CHF or for AMI. This targeting method, based on binomial probability, is described in more detail below.
2. NOS 30-day postadmission death targeting, applied to the same populations. This targeting method was the same as for inpatient targeting, but it was applied to death within 30 days of admission to the hospital, rather than death in the hospital.
3. HCFA 30-day death targeting, applied to the last discharge for each patient, during calendar year 1986 or during calendar year 1987, of all Medicare patients (including those under 65), from U.S. acute care hospitals, for severe chronic heart disease or for severe acute heart disease. This targeting method, based on residuals from logistic regression models, is described in more detail below.

TARGETING FOR THE NONINTRUSIVE OUTCOMES STUDY

Inpatient Deaths

We obtained information on all hospital stays for Medicare beneficiaries from HCFA's Bill Record File for all admissions occurring between October 1, 1983, and September 30, 1984. We obtained additional information on hospitals from HCFA's Provider of Service File. To make the data as comparable as possible across hospitals, we (1) excluded from the analysis all Medicare beneficiaries under the age of 65 (those eligible to receive Medicare benefits because of various disabilities, including chronic renal disease); (2) excluded data from long term care hospitals, psychiatric facilities, hospices, and rehabilitation

hospitals; (3) excluded interim bills; (4) edited the data to include only one complete record for each hospital stay; and (5) counted transfers from one acute care hospital to another as live discharges from the first hospital.

We defined congestive heart failure as DRG 127 and acute myocardial infarction as DRGs 121, 122, 123, and 115. For both conditions, we also required appropriate ICD-9 codes for the principal diagnosis; the specific codes are 398.91, 402.11, 402.91, 428.0, 428.1, 428.9, or 785.51 for congestive heart failure, and 410.0 through 410.9 for acute myocardial infarction.

For each hospital, we calculated d = the death rate it would have experienced if its CHF or AMI patients had died at nationwide average rates for each condition for each of 20 age-sex-race cells. We then calculated the binomial probability that a hospital whose n patients each had a true probability of dying d, would have as many deaths m as it actually did, $p(d,n,m)$. Hospitals with less than a 0.05 probability of having as many deaths as they did, $p(d,n,m) < 0.05$, were called targeted; all others were untargeted. (For additional details on NOS targeting for CHF, AMI, and 46 other frequently occurring specific conditions, together with deaths for all admissions, see Chassin et al., 1989.)

Some summary statistics for inpatient targeting nationwide are at the top of Table A.1. Of 5787 acute care hospitals that treated Medicare elderly CHF patients, 7.2 percent were targeted at the 5 percent level. Hospitals with more such patients were more likely to be targeted than were smaller hospitals, because the binomial probability test has more power when n is larger. The overall CHF inpatient death rate in the administrative data was 9.7 percent; targeted hospitals had a higher rate, 16.7 percent.

Using AMI inpatient deaths, 7.7 percent of hospitals were targeted. Again, targeted hospitals were on average larger than untargeted hospitals. The AMI inpatient death rate was 21.2 percent overall, and 33.8 percent in targeted hospitals.

Deaths Within 30 Days of Admission

To target high death rate hospitals using deaths within 30 days of admission, we obtained information on dates of out-of-hospital death from HCFA's Health Insurance Master File. The file we used included records only for people that the Social Security Administration listed as deceased. We linked these data to the Bill Record File using social security number (SSN) and gender. We counted the patient as being dead within 30 days of admission if either he was discharged dead within 30 days of admission according to the Bill Record File, or he

Table A.1

NATIONWIDE SUMMARY STATISTICS FOR VARIOUS TARGETING METHODS

	Targeting Method	Number of Hospitals	Percent of Hospitals	Average Number of Patients	Death Rate (%)
(1)	NOS inpatient, CHF, 1984	5787	100.0	79.7	9.7
	Untargeted	5372	92.8	75.3	8.8
	Targeted	415	7.2	136.5	16.7
(2)	NOS inpatient, AMI, 1984	5702	100.0	47.2	21.2
	Untargeted	5262	92.3	45.8	19.6
	Targeted	440	7.7	64.3	33.8
(3)	NOS 30-day, CHF, 1984	5787	100.0	79.7	14.1
	Untargeted	5426	93.8	78.0	13.4
	Targeted	361	6.2	104.0	22.3
(4)	NOS 30-day, AMI, 1984	5702	100.0	47.2	25.4
	Untargeted	5270	92.4	47.0	24.2
	Targeted	432	7.6	49.7	39.6
(5)	HCFA 30-day, chronic, 1986	5657	100.0	54.2	22.5
	Untargeted	5488	97.0	54.1	22.1
	Targeted	169	3.0	55.4	36.2
(6)	HCFA 30-day, acute, 1986	5580	100.0	45.2	38.1
	Untargeted	5060	90.7	46.4	36.8
	Targeted	520	9.3	34.5	56.0
(7)	HCFA 30-day, chronic, 1987	5645	100.0	55.2	22.0
	Untargeted	5464	96.8	55.1	21.5
	Targeted	181	3.2	58.5	34.9
(8)	HCFA 30-day, acute, 1987	5560	100.0	44.4	37.3
	Untargeted	4944	88.9	46.1	35.8
	Targeted	616	11.1	30.2	55.9

SOURCE: Calculated from administrative data.

NOTE: Death rates are for groups of hospitals (all, untargeted, and targeted), that is, they are weighted average death rates for hospitals in each group.

matched a death date withing 30 days of admission in the Health Insurance Master File.

We did not investigate the validity of the out-of-hospital death match in any detail, but some information is available from a similar effort by the RAND PPS study. PPS attempted to match all of their medical record abstracts to a health insurance file that contained information for all beneficiaries, dead or alive. They used SSN, beneficiary identification code, name, and date of birth. Using both computer and supplemental hand matching, they were able to match 92 percent of the abstracts to the health insurance file.

Our match shows that 14.1 percent of CHF patients died within 30 days of admission, compared with 9.7 percent who died in the hospital (Table A.1). The corresponding figures for AMI are 25.4 and 21.2. Assuming that we matched 92 percent of deaths within 30 days of admission, these figures understate 30-day deaths only slightly. For CHF, the inpatient death rate is unaffected because it is based on the Bill Record File itself, but the increment in deaths following discharge might have been (14.1 − 9.7)/0.92 = 4.8, so the overall 30-day death rate might have been 14.5 instead of 14.1. A similar calculation for AMI suggests that the 30-day death rate might be 25.8 rather than 25.4.

We calculated the expected 30-day death rate for each hospital and targeted if the binomial probability of having as high a 30-day death rate as actually observed was less than 0.05. Summary statistics for NOS 30-day targeting appear in Table A.1. For CHF, 6.2 percent of hospitals were targeted, and for AMI, 7.6 percent. Targeted hospitals were larger than untargeted for CHF, as expected, but (unexpectedly) about the same size for AMI. Deaths within 30 days of admission were about 40 percent higher than inpatient deaths for CHF, and about 20 percent higher for AMI.

HCFA TARGETING

Our characterization of HCFA targeting is based on information in their 1988 data release, "Medicare Hospital Mortality Information," covering calendar years 1987 and 1986. This document states that

> The principal source of data for the analysis was the HCFA Medicare Provider Analysis and Review (MEDPAR) file which contains information about each Medicare hospital discharge. Information about beneficiaries, including date of death, was obtained from the Social Security Administration. The analyses were performed by means of logistic regression, with patients grouped into 17 distinct diagnostic categories [defined by ICD-9 code and including severe chronic heart disease and severe acute heart disease.] . . . [T]he risk factors (covariates) included in the regression analyses within each diagnostic category were age, sex, significant comorbidities, number of admissions by risk category within 6 months, and status as a transfer patient.

Severe chronic heart disease is defined by HCFA more broadly than is NOS CHF, to include, for example, chronic pulmonary heart disease and certain cardiomyopathies in addition to CHF. Severe acute heart disease is also broader than NOS AMI, including for example acute pulmonary heart disease, bacterial endocarditis, cardiopulmonary

arrest, and ruptured thoracic or abdominal aneurysms in addition to AMI. See Bowen and Roper, 1988, for a complete list of the ICD-9 codes that define the HCFA conditions.)

The HCFA data release includes for each hospital for each condition the number of cases, the actual death rate in percent, and lower and upper bounds on the predicted death rate in percent. The size of the range between lower and upper bounds differs to account for the differences in variability given the number of cases. We treat the range here as though it were calculated from a 95 percent confidence interval around the log odds predicted by the logit regression. (In fact, the HCFA method differed from year to year, and particularly for the analysis of 1987 data, it was more complicated than our simplified treatment here would indicate. See Bowen and Roper, 1988, for details.) Thus a hospital whose actual death rate exceeded the upper bound on predicted death rate is targeted at the 2.5 percent level, in the same sense that the NOS hospitals are targeted at the 5 percent level. That is, our stylization of HCFA's test is a two-tailed test at the 0.05 level, whereas the NOS test is one-tailed at the 0.05 level.

To increase direct comparability with NOS targeting, we calculated from the HCFA data the implied probability that a hospital would have as high a death rate as observed, and redesignated the hospital as targeted according to the HCFA method if that probability were less than 0.05 (one tailed). To do so, we (1) reset rates of 0 or 100 percent to 0.1 or 99.9 percent, (2) transformed all rates to logits (log odds), (3) calculated the mean predicted logit as the mean of the upper and lower bound logits, (4) calculated the standard deviation of the predicted logit as the range between the upper and lower bound logits divided by 2×1.96, (5) calculated the z score as actual logit minus mean predicted logit divided by standard deviation of predicted logit, and (6) calculated probability as 1 minus the cumulative normal probability of z.

Summary statistics for HCFA targeting at the 5 percent level, for two conditions in each of two years, are in Table A.1. Curiously, HCFA targets only about 3 percent of hospitals for chronic heart disease; this seems to imply that there is *less* variation in death rates for this condition than one would expect on the basis of chance alone. In contrast, they target as many as 11 percent of hospitals for acute heart disease. Targeted and untargeted hospitals are on average the same size for chronic heart disease (that is, targeted and untargeted hospitals treated the same number of CHF patients on average). For acute heart disease, the HCFA method tends to target smaller hospitals. This is in contrast to the NOS method, which as noted above has more power to target larger hospitals.

HCFA death rates are about 50 percent higher than the corresponding NOS death rates. Some of the difference is accounted for by HCFA's use of only the last discharge of the year (omitting previous, necessarily live, discharges), whereas NOS uses all admissions. For the great bulk of the discharges that HCFA ignores, the patient would still be alive 30 days after the corresponding admission. Some of the difference may also be accounted for by differences in the HCFA and NOS definitions of the conditions. For example, acute severe heart disease includes cardiac arrest (ICD-9-CM code 427.5), a very high death rate diagnosis that is not included in acute myocardial infarction (ICD-9-CM codes 410.0 through 410.9).

CORRELATIONS AMONG TARGETING METHODS

Table A.2 shows Pearson correlations among the probabilities that a hospital would have as many deaths as actually observed, calculated for each targeting method, condition, and year. The correlations are all positive. Certain correlations of particular interest are enclosed in brackets [], braces {}, or parentheses (). Additional results for the two NOS targeting methods plus two others, and for a score of other conditions, are in Chassin et al. (1989).

Table A.2

NATIONWIDE CORRELATIONS AMONG PROBABILITIES OF HAVING AS MANY DEATHS AS ACTUALLY EXPERIENCED, FOR VARIOUS TARGETING METHODS
(N = 5348 hospitals)

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1)	NOS inpatient, CHF, 1984	1.00							
(2)	NOS inpatient, AMI, 1984	[0.26]	1.00						
(3)	NOS 30-day, CHF, 1984	(0.70)	0.11	1.00					
(4)	NOS 30-day, AMI, 1984	0.12	(0.80)	[0.12]	1.00				
(5)	HCFA 30-day, chronic, 1986	0.16	0.12	{0.19}	0.12	1.00			
(6)	HCFA 30-day, acute, 1986	0.05	0.14	0.07	{0.16}	[0.12]	1.00		
(7)	HCFA 30-day, chronic, 1987	0.13	0.13	{0.16}	0.13	{0.23}	0.12	1.00	
(8)	HCFA 30-day, acute, 1987	0.04	0.13	0.05	{0.15}	0.12	{0.29}	[0.11]	1.00

SOURCE: Calculated from administrative data.

NOTE: Numbers in parentheses denote correlations across death measures (same year and same condition). Numbers in brackets denote correlations across conditions (same death measure and same year). Numbers in braces denote correlations across years (same death measure and same condition).

The correlations between NOS inpatient probabilities and 30-day probabilities (enclosed in parentheses) are quite high: 0.70 for CHF and 0.80 for AMI. These correlations correspond to a substantial overlap in hospitals targeted at the 5 percent level using NOS inpatient and 30-day methods. For AMI, a cross tabulation of hospitals targeted by the two methods shows that of the 440 hospitals targeted for inpatient deaths, 60 percent are also targeted for 30-day deaths. Of 432 that are 30-day targeted, 61 percent are also inpatient targeted.

The correlations between the probabilities for different conditions, calculated using the same targeting method applied to the same year (enclosed in brackets) range from 0.10 up to 0.26. The overlap between hospitals targeted for different conditions is fairly small. A cross tabulation of hospitals targeted at the 5 percent level by HCFA during 1986 for chronic and acute heart disease, for example, shows that of the 171 hospitals targeted for chronic heart disease, 15 percent are also targeted for acute heart disease. Of 523 targeted for chronic heart disease, only 5 percent were also targeted for acute heart disease.

The correlations between years for the same condition and same targeting method (enclosed in braces) tend to be somewhat higher, 0.23 for chronic heart disease in 1986 and 1987, and 0.29 for acute heart disease in 1986 and 1987. A cross tabulation for acute heart disease shows that 25 percent of the 523 hospitals targeted in 1986 are also targeted in 1987, and 21 percent of the 625 hospitals targeted in 1987 were also targeted in 1986.

Disregarding the differences in the way the conditions are defined, one can also compare NOS 30-day targeting of CHF in 1984 with HCFA targeting of chronic heart disease in 1986 and 1987, and similarly for NOS AMI and HCFA acute heart disease. The correlations here range from 0.15 to 0.19. One would expect them to be lower than the 1986/1987 correlations within disease both because of differences in disease definitions and because of the longer time interval between observations. The pattern of correlations across time is consistent with the existence of persistent hospital effects together with an autocorrelated random error effect. In Table A.3, we present correlations of the logistic transformations of the probabilities, rather than the probabilities themselves. The transformation makes very little difference in the correlations.

Table A.3

NATIONWIDE CORRELATIONS AMONG LOGISTIC TRANSFORMATIONS OF PROBABILITIES OF HAVING AS MANY DEATHS AS ACTUALLY EXPERIENCED, FOR VARIOUS TARGETING METHODS
(N = 5348 hospitals)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1) NOS inpatient, CHF, 1984	1.00							
(2) NOS inpatient, AMI, 1984	[0.25]	1.00						
(3) NOS 30-day, CHF, 1984	(0.69)	0.09	1.00					
(4) NOS 30-day, AMI, 1984	0.10	(0.80)	[0.12]	1.00				
(5) HCFA 30-day, chronic, 1986	0.11	0.07	{0.14}	0.08	1.00			
(6) HCFA 30-day, acute, 1986	0.04	0.12	0.06	{0.16}	{0.11}	1.00		
(7) HCFA 30-day, chronic, 1987	0.11	0.11	{0.14}	0.12	{0.19}	0.11	1.00	
(8) HCFA 30-day, acute, 1987	0.03	0.11	0.03	{0.14}	0.07	{0.27}	{0.12}	1.00

SOURCE: Calculated from administrative data.

NOTE: Numbers in parentheses denote correlations across death measures (same year and same condition). Numbers in brackets denote correlations across conditions (same death measure and same year). Numbers in braces denote correlations across years (same death measure and same condition).

TARGETING AND HOSPITAL CHARACTERISTICS

We have already noted that larger hospitals are more likely to be targeted by NOS, and smaller hospitals are more likely to be targeted by HCFA for acute heart disease. In this subsection, we check for relationships between targeting and other hospital characteristics in a multivariate framework. Most of the hospital characteristics come from the HCFA Hospital Record File They are: number of certified beds, type of controlling organization (church, proprietary, government, or other), extent of medical school affiliation (major, limited, graduate, or none), presence or absence of a residency program, size category of the hospital's standard metropolitan statistical area (down to no SMSA at all), and region of the country.

A natural approach would be to run logistic regressions with the dependent variable being whether or not a hospital was targeted and the independent variables being hospital characteristics. The problem with this approach is that by categorizing hospitals as targeted or not depending on whether p is less than or greater than 0.05, much of the information in the p scale is thrown away. We chose to retain that information and do ordinary least squares regressions of p, the probability of having as many deaths as observed, on hospital characteristics.

This provides estimates that are similar to, but more precise than, the logistic regression estimates.

The results are in Table A.4. A negative coefficient means that an increase in the variable decreases the probability of so many deaths.

Table A.4

NATIONWIDE REGRESSIONS OF PROBABILITY (%) OF HAVING AS MANY DEATHS AS OBSERVED ON HOSPITAL CHARACTERISTICS
(N = 5348 hospitals)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Beds	−4.11	−2.69	−2.43	−0.66	−1.04	1.06	−0.92	1.76
	(−12.1)	(−7.7)	(−7.3)	(−1.9)	(−3.2)	(3.1)	(−2.8)	(5.1)
Church	−1.27	−3.59	−1.12	−3.09	−0.66	−1.36	−1.91	−3.24
	(−0.9)	(−2.5)	(−0.8)	(−2.2)	(−0.5)	(−1.0)	(−1.4)	(−2.3)
Propri	−0.64	−1.24	−1.51	−3.90	0.78	−5.46	−0.20	−5.46
	(−0.5)	(−0.8)	(−1.1)	(−2.7)	(0.6)	(−3.8)	(−0.1)	(−3.7)
Gvt	0.57	−1.07	−3.68	−4.03	0.51	−6.59	−1.58	−4.87
	(0.6)	(−1.0)	(−3.6)	(−3.8)	(0.5)	(−6.3)	(−1.6)	(−4.5)
Major	1.94	−3.72	5.91	−3.21	8.43	−1.09	6.87	−2.60
	(0.8)	(−1.5)	(2.5)	(−1.3)	(3.6)	(−0.4)	(2.9)	(−1.1)
Limited	−4.42	−3.93	0.64	−3.17	0.44	−0.44	−0.74	−3.35
	(−2.5)	(−2.2)	(0.4)	(−1.8)	(0.3)	(−0.2)	(−0.4)	(−1.8)
Graduate	−6.62	−1.99	−5.47	−1.17	−3.08	−4.76	−3.28	−7.03
	(−1.8)	(−0.5)	(−1.5)	(−0.3)	(−0.9)	(−1.3)	(−0.9)	(−1.9)
Res_pgm	2.09	4.57	1.06	4.33	4.46	3.54	6.53	5.33
	(1.0)	(2.1)	(0.5)	(2.0)	(2.2)	(1.7)	(3.2)	(2.5)
Rural	3.51	−0.84	1.55	−3.90	0.59	2.47	1.35	1.15
	(3.4)	(−0.8)	(1.5)	(−3.7)	(0.6)	(2.4)	(1.3)	(1.1)
Constant	67.16	63.96	61.67	59.82	53.90	47.98	54.25	46.65
	(60.7)	(55.7)	(56.5)	(53.4)	(50.4)	(42.7)	(50.1)	(40.7)
R-square	0.06	0.02	0.02	0.01	0.01	0.02	0.01	0.02
Observations	5785	5698	5785	5698	5566	5501	5541	5462

SOURCE: Calculated from administrative data.

NOTES: Dependent variable is probability for targeting method. The targeting methods are: (1) NOS inpatient, CHF, 1984; (2) NOS inpatient, AMI, 1984; (3) NOS 30-day, CHF, 1984; (4) NOS 30-day, AMI, 1984; (5) HCFA 30-day, severe chronic heart disease, 1986; (6) HCFA 30-day, severe acute heart disease, 1986; (7) HCFA 30-day, severe chronic heart disease, 1987; and (8) HCFA 30-day, severe acute heart disease, 1987. Independent variables are: beds—number of certified beds; church—hospital is operated by a religious organization; propri—hospital is operated by a for profit organization; gvt—hospital is operated by a government organization; major—hospital has a major affiliation with a medical school; limited—hospital has a limited affiliation with a medical school; graduate—hospital has an affiliation with a graduate medical program; res_pgm—hospital has a residency program; and rural—hospital is located in a rural area. The results shown are fairly stable even when 10 regional dummies and seven SMSA size dummies are included.

Thus a negative sign means that targeting is *more* likely (minus is bad). For example, the negative coefficients for beds for the NOS targeting methods confirm the univariate result that larger hospitals are more likely to be targeted, and the positive coefficients on beds for HCFA acute targeting confirm that smaller hospitals are more likely to be targeted. Interestingly, although there was no univariate relationship between size and targeting for HCFA chronic heart disease, there is one in the multivariate regressions: Larger hospitals are more likely to be targeted.

The results are otherwise fairly consistent across targeting methods. The following are generally more likely to be targeted: church run hospitals, proprietary hospitals, government operated hospitals, and hospitals with only limited or graduate medical school affiliations. The following are generally less likely to be targeted: hospitals with a major medical school affiliation and hospitals with a residency program. The effect of rural location is mixed. Although many of the coefficients are statistically significant, none of the equations has much explanatory power. R-squared is usually only 0.01 or 0.02.

In Table A.4 equations, the only city size variable is rural or not, and the region of the country is not included. Unreported regressions that also include seven dummy variables for city size dummies and 10 dummies for region of the country do not change the results in Table A.4 very much.

Appendix B

SAMPLING IN FOUR STATES

OVERVIEW

The main objectives of the NOS were (1) to determine if hospitals with high inpatient death rates provide lower quality care than do hospitals with lower death rates, and (2) to determine how inpatient death is related to severity of illness and quality of care.

We investigated these questions for two medical conditions: CHF and AMI, using detailed abstracts of a sample of medical records for Medicare elderly patients (age 65 or older) discharged from the hospital during fiscal year 1984.

For each condition separately, we divided hospitals into two categories: those with less than a 0.05 probability of having as many inpatient deaths as they did, after adjusting for the age, sex, and race distribution of their patients ("targeted" hospitals), and all others ("untargeted" hospitals). (See Chassin et al., 1989, for details.) Framed in these terms, question (1) asks whether targeted hospitals provide lower quality care than do untargeted hospitals.

Patients may be categorized as those who died during their hospital stay and those who did not. Then question (2) asks how well can one explain outcomes for individual patients (dead or alive) on the basis of severity and quality measures.

To answer question (1) most precisely, we would want equal-sized samples of patients in targeted and untargeted hospitals. To answer question (2) most precisely, we would want equal-sized samples of patients discharged dead and patients discharged alive. We accommodated both goals by drawing equal numbers of sample patients in the four cells defined by targeted compared with untargeted hospitals and dead compared with alive discharges.

For logistic reasons, we confined the sample to four states (California, Illinois, Minnesota, and New York). These states have 20 percent of U.S. hospitals and 22 percent of Medicare hospitalizations. Some statistics for hospitals and patients in the four states, comparable to some of those reported in Appendix A for the nation as a whole, are given in the next subsection.

Power calculations suggested that a sample of 350 patients in each of the four cells for each of the two conditions would be adequate. We

drew a systematic stratified random sample of discharges in the following manner. Our sample frame consisted of Medicare claims records arranged into eight lists. There was a separate list for each condition for each of the four targeted/untargeted, dead/alive cells. We sorted each list by state and hospital; within each hospital, we listed patients in random order. For each list separately, we calculated a sampling interval, I, as the number of patients on the list divided by 400. We then selected every Ith patient on the list for inclusion in the sample. That interval yielded a sample of 400 from each list, a little over 350 to allow for some unobtainable records.

This systematic random sampling procedure assured that the sample within each of the eight condition/targeting/death cells was representative of all hospitalizations, hospitals, and states in that cell. At the same time, by oversampling targeted hospitals we increased our ability to discriminate between the quality of care provided by targeted compared with untargeted hospitals, and by oversampling patients discharged dead we increased our ability to estimate the effects of severity of illness and quality of care on the probability that a patient would die in the hospital.

This section demonstrates and quantifies the benefits of our sample relative to a purely proportional sample, first for the two main study issues for which the sampling plan was actually designed, and then for an alternative way of targeting hospitals that became important during the course of this study. The alternative targeting method targets hospitals that have unexpectedly high death rates within 30 days of admission, whether the patient was still in the hospital at that time or not. It is not immediately clear that a sampling plan designed to investigate inpatient targeting will work well for 30-day targeting too, but that does turn out to be true.

In the field, we found that substantial numbers of claims records were misleading as to the reason for hospitalization; many claims involving CHF (or AMI) turned out to represent hospitalizations for other reasons, as determined by examining the medical record. For this and other reasons, the realized sample was smaller than the hoped for 350 hospitalizations in each of eight cells. The comparison of our sampling plan to a purely proportional sampling plan is based on the smaller realized sample (and the correspondingly smaller population of actual CHFs (or AMIs)). Before turning to the comparison of sampling methods, then, we describe sample attrition in the second following subsection.

THE FOUR STATES ARE SIMILAR TO THE NATION AS A WHOLE

Table B.1 replicates for the four sample states the summary statistics that Table A.1 shows for the nation as a whole. Death rates in the

Table B.1

SUMMARY STATISTICS IN FOUR SAMPLE STATES FOR VARIOUS TARGETING METHODS

	Targeting Method	Number of Hospitals	Percent of Hospitals	Average Number of Patients	Death Rate (%)
(1)	NOS inpatient, CHF, 1984	1137	100.0	94.3	11.1
	Untargeted	992	87.2	84.6	9.4
	Targeted	145	12.8	160.1	17.3
(2)	NOS inpatient, AMI, 1984	1121	100.0	55.9	21.7
	Untargeted	1017	90.7	54.0	20.0
	Targeted	104	9.3	74.7	33.7
(3)	NOS 30-day, CHF, 1984	1137	100.0	94.3	14.5
	Untargeted	1063	93.5	91.3	13.7
	Targeted	74	6.5	136.7	21.6
(4)	NOS 30-day, AMI, 1984	1121	100.0	55.9	24.6
	Untargeted	1049	93.6	55.8	23.6
	Targeted	72	6.4	56.6	39.4
(5)	HCFA 30-day, chronic, 1986	1103	100.0	62.8	22.4
	Untargeted	1076	97.6	63.0	22.1
	Targeted	27	2.4	52.9	37.9
(6)	HCFA 30-day, acute, 1986	1090	100.0	48.8	38.1
	Untargeted	967	88.7	52.2	36.4
	Targeted	123	11.3	38.8	56.7
(7)	HCFA 30-day, chronic, 1987	1106	100.0	63.7	21.8
	Untargeted	1080	97.6	63.9	21.5
	Targeted	26	2.4	54.7	36.8
(8)	HCFA 30-day, acute, 1987	1083	100.0	49.4	37.3
	Untargeted	962	88.8	51.5	35.7
	Targeted	121	11.2	32.2	56.8

SOURCE: Calculated from administrative data.

NOTE: Death rates are for groups of hospitals (all, untargeted, and targeted), that is, they are weighted average death rates for hospitals in each group.

four states, and the spread between death rates in targeted and untargeted hospitals, are similar to those nationwide. The average hospital in the four states is somewhat larger (that is, it treats more CHF and AMI patients) than does the nationwide average hospital. Also, a somewhat larger fraction of hospitals are targeted by some of the targeting methods in the four states relative to the nation as a whole.

Table B.2 shows correlations among the various targeting methods in the four states, corresponding to the nationwide correlations shown in Table A.2. Again, the results in the four states are quite similar to those nationwide.

SAMPLE ATTRITION

Table 3 described the reasons for sample attrition and showed how many usable sample records were obtained after attrition. Most of the attrition was for "unavoidable" reasons. For example, upon examination of sampled CHF records, it turned out that 248 out of 1600 were coding errors (Table 3); CHF was not the true reason for these 248 hospitalizations even though it was coded as the principal diagnosis in

Table B.2

CORRELATIONS IN FOUR SAMPLE STATES AMONG PROBABILITIES OF HAVING AS MANY DEATHS AS ACTUALLY EXPERIENCED, FOR VARIOUS TARGETING METHODS
(N = 1052 hospitals)

		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1)	NOS inpatient, CHF, 1984	1.00							
(2)	NOS inpatient, AMI, 1984	[0.24]	1.00						
(3)	NOS 30-day, CHF, 1984	(0.67)	0.11	1.00					
(4)	NOS 30-day, AMI, 1984	0.12	(0.85)	[0.11]	1.00				
(5)	HCFA 30-day, chronic, 1986	0.13	0.08	{0.18}	0.07	1.00			
(6)	HCFA 30-day, acute, 1986	0.04	0.14	0.08	{0.19}	{0.11}	1.00		
(7)	HCFA 30-day, chronic, 1987	0.16	0.11	{0.20}	0.10	{0.24}	0.14	1.00	
(8)	HCFA 30-day, acute, 1987	0.03	0.17	0.03	{0.20}	0.11	{0.35}	[0.10]	1.00

SOURCE: Calculated from administrative data.

NOTE: Numbers in parentheses denote correlations across death measures (same year and same condition). Numbers in brackets denote correlations across conditions (same death measure and same year). Numbers in braces denote correlations across years (same death measure and same condition).

the claims data. This means that the apparent population of CHF hospitalizations based on claims records is wrong; many were not really CHF hospitalizations. We calculated adjusted population values in each of the four targeted/untargeted dead/alive cells by scaling down the number of claims records that appeared to represent CHF hospitalizations in proportion to the sample attrition in each cell. It is these adjusted population counts that are shown for inpatient targeted and untargeted hospitals in Table 3.

Table B.3 shows more information on realized sample and adjusted population counts. The additional detail distinguishes between dead and alive patients as well as targeted and untargeted hospitals. It also shows sample and adjusted population counts for two targeting methods in addition to targeting based on inpatient deaths: targeting based on deaths within 30 days of admission in our 1984 data, and targeting by HCFA using 1986 data. (See Appendix A for a discussion of the various targeting methods.) "Dead" for inpatient targeting means dead at discharge from the hospital; for 30-day and HCFA targeting, it means dead within 30 days of admission to the hospital. For 30-day and HCFA targeting, the population counts were adjusted in proportion to sample attrition in each of 16 cells that result from crossing the four dead/alive targeted/untargeted inpatient categories with the corresponding four categories for the alternative targeting method.

Our sample was designed to oversample patients discharged from hospitals targeted using inpatient deaths; it also turned out to oversample patients from 30-day targeted hospitals. This is because of the high correlation between hospital probabilities of having as many inpatient deaths as observed with probabilities of having as many 30-day deaths as observed (Appendix A). It also oversampled patients from HCFA targeted hospitals for a similar reason, but even so the sample from HCFA targeted hospitals is quite thin, especially for CHF patients (only 26 of them).

The population death rates after adjustment to reflect sample attrition tend to be lower than the death rates based on administrative data before adjustment (Table B.1). This probably reflects the fact that many of the coding errors occurred when cause of death was mistakenly coded as reason for admission.

Population death rates are about the same in 1984 for hospitals targeted and those not targeted by HCFA using 1986 data.

It is important to keep in mind that we sampled patients, not hospitals. The first study question concerns differences between care received by patients treated in all targeted hospitals as a group versus those treated in untargeted hospitals as a group. The sample is not designed to say anything about care provided by individual hospitals.

Table B.3

ADJUSTED POPULATION AND SAMPLE COUNTS BY SAMPLING CATEGORY AFTER SAMPLE ATTRITION

	Inpatient Targeting		30-Day Targeting		HCFA Targeting	
	Untargeted Hospitals	Targeted Hospitals	Untargeted Hospitals	Targeted Hospitals	Untargeted Hospitals	Targeted Hospitals
CHF Patients						
Hospitals						
Four states	992	145	1063	74	1076	27
Sample	440	131	515	56	545	15
Patients in four study states						
Alive	60482	13930	65422	6036	69048	1132
Dead	5220	2534	9422	1286	10366	169
Total	65702	16464	74844	7322	79413	1301
Death rate	7.94	15.39	12.59	17.56	13.05	12.97
Patients in sample						
Alive	318	290	526	109	610	13
Dead	265	253	402	89	470	13
Total	583	543	928	198	1080	26
AMI Patients						
Hospitals						
Four states	1017	104	1049	72	968	153
Sample	453	92	492	53	525	11
Patients in four study states						
Alive	34124	3667	34511	1974	32938	2799
Dead	8540	1590	10416	1020	10148	915
Total	42664	5257	44927	2994	43087	3714
Death rate	20.02	30.25	23.18	34.07	23.55	24.63
Patients in sample						
Alive	311	285	464	129	508	73
Dead	311	243	429	128	476	68
Total	622	528	893	257	984	141

Table B.4 shows that the majority of sampled patients come from hospitals with only one or two patients in the sample for each condition. Only four hospitals have 10 or more CHF patients in the sample, and only 14 have 10 or more AMI patients in the sample.

COMPARATIVE SAMPLE PERFORMANCE FOR MAIN STUDY QUESTIONS

The Framework for the Comparison

We investigate the performance of our sample by looking at how well it answers the following two questions. The questions are meant to be representative of the many detailed questions that can be answered to accomplish the two main study objectives set out at the start of this appendix: (1) What is the difference between the average quality of care in targeted and untargeted hospitals? (2) What is the difference between the average quality of care for patients discharged dead and those discharged alive?

Here quality of care is measured by the overall process score developed by the PPS study. This process score is scaled to have mean 0.0 and standard deviation 1.0 in the PPS data; it has about the same distribution in the NOS data.

Table B.4

SAMPLED PATIENTS BY NUMBER OF SAMPLED PATIENTS PER HOSPITAL

	CHF Patients			AMI Patients		
Sampled Patients	Untargeted Hospitals	Targeted Hospitals	Total	Untargeted Hospitals	Targeted Hospitals	Total
1	311	19	330	304	13	317
2	230	42	272	268	24	292
3	42	63	105	33	27	60
4	0	76	76	12	36	48
5	0	80	80	5	35	40
6	0	90	90	0	30	30
7	0	35	35	0	91	91
8	0	48	48	0	48	48
9	0	45	45	0	36	36
10	0	10	10	0	40	40
11	0	22	22	0	11	11
12	0	0	0	0	24	24
13	0	13	13	0	0	0
14	0	0	0	0	0	0
15	0	0	0	0	45	45
16	0	0	0	0	16	16
17	0	0	0	0	34	34
18	0	0	0	0	18	18
Total	583	543	1126	622	528	1150

We explicitly compare three ways to answer those two questions:

1. Unweighted estimates based on a hypothetical proportional sample, in which each of the four targeted/untargeted dead/alive cells is represented in proportion to the population number of hospitalizations in each cell.
2. Weighted estimates based on our actual sample, which oversamples hospitalizations in targeted hospitals and those that end in death. Population weights are used to get unbiased estimates for the two contrasts of interest.
3. Unweighted estimates using our actual sample. These will in general be biased, but they may nonetheless be better than the unbiased weighted estimates in the sense of having lower expected root mean squared errors (RMSE).

In fact, we will be interested in many things in addition to these two simple contrasts between unconditional means of overall process scores in two groups of patients (those in targeted compared with untargeted hospitals, those who died compared with those who did not). For example, we will compare process subscales, severity of illness, and other things in addition to the overall process. And we will compare conditional means in a regression framework, controlling for the effect of other variables in addition to the two-way categorization into targeted compared with untargeted or dead compared with alive. And we will be interested in regression coefficients that measure the effect of one variable on another after controlling for covariates, in addition to simple means.

The relative performance of the three methods should be the same in answering these more complex questions, subject to certain conditions discussed at the end of this subsection.

The Detailed Numbers

Table B.5 describes the population and the actual sample, together with a hypothetical proportional sample. The raw weights, which show the number of population hospitalizations represented by each sample observation, reflect the oversampling of targeted hospitals and dead discharges. The population means postulated for the four sampling cells for the purposes of this comparison are equal to the observed sample means in each cell (which are unbiased estimates of the population values). The means for the marginal "total" row and column are calculated from the four cell means using population weights. The standard deviation is assumed equal to 1.0 in all cells for computational convenience and because it is nearly true. Table B.6 shows the details of the

Table B.5

POPULATION AND SAMPLES FOR SAMPLING COMPARISON: INPATIENT DEATHS AND TARGETING[a]

	CHF Patients			AMI Patients		
	Untargeted Hospitals	Targeted Hospitals	Total	Untargeted Hospitals	Targeted Hospitals	Total
Population						
Alive	60482	13930	74412	34124	3667	37791
Dead	5220	2534	7754	8540	1590	10130
Total	65702	16464	82166	42664	5257	47921
Actual sample						
Alive	318	290	608	311	285	596
Dead	265	253	518	311	243	554
Total	583	543	1126	622	528	1150
Raw weights						
Alive	190.2	48.0	122.4	109.7	12.9	63.4
Dead	19.7	10.0	15.0	27.5	6.5	18.3
Total	112.7	30.3	73.0	68.6	10.0	41.7
Proportional sample						
Alive	829	191	1020	819	88	907
Dead	72	35	106	205	38	243
Total	900	226	1126	1024	126	1150
Population means						
Alive	0.100	0.220	0.122	0.315	0.347	0.318
Dead	−0.270	−0.140	−0.228	0.052	0.047	0.051
Total	0.071	0.165	0.089	0.262	0.256	0.262

[a]Standard deviation is assumed equal to 1.00 in all cells.

results for the three estimation methods. For method (1), unweighted sample means in each of the cells (including the "total" row and column) are unbiased estimates of the population means. The standard errors are just the inverse of the square root of the (proportional) sample sizes. The proportional sample provides the minimum variance estimate of the overall population mean, but the estimates for the less populous cells are less reliable. In particular, cells for targeted hospitals and dead discharges have relatively high standard errors, and that is a clear disadvantage for the comparisons that we are making here.

Method (2), weighted estimates from the actual sample, also provides unbiased estimates. For the four sampling cells, standard errors are (nearly) equal because sample sizes are (nearly) equal. The marginal standard errors, calculated as the square root of the population

Table B.6

ESTIMATES FOR SAMPLE COMPARISON: INPATIENT DEATHS AND TARGETING

	CHF Patients			AMI Patients		
	Untargeted Hospitals	Targeted Hospitals	Total	Untargeted Hospitals	Targeted Hospitals	Total
(1) Unweighted Estimates from Proportional Sample Are Unbiased						
Standard errors						
Alive	0.035	0.072	0.031	0.035	0.107	0.033
Dead	0.118	0.170	0.097	0.070	0.162	0.064
Total	0.033	0.067	0.030	0.031	0.089	0.029
(2) Weighted Estimates from Actual Sample Are Unbiased						
Standard errors						
Alive	0.056	0.059	0.047	0.057	0.059	0.052
Dead	0.061	0.063	0.046	0.057	0.064	0.049
Total	0.052	0.051	0.043	0.047	0.046	0.042
(3) Unweighted Estimates from Actual Sample Are Biased but May Have Smaller RMSE						
Estimated means						
Alive	0.100	0.220	0.157	0.315	0.347	0.330
Dead	−0.270	−0.140	−0.207	0.052	0.047	0.050
Total	−0.068	0.052	−0.010	0.184	0.209	0.195
Bias						
Alive	0.000	0.000	0.035	0.000	0.000	0.012
Dead	0.000	0.000	0.021	0.000	0.000	−0.001
Total	−0.139	−0.112	−0.100	−0.079	−0.047	−0.067
Standard errors						
Alive	0.056	0.059	0.041	0.057	0.059	0.041
Dead	0.061	0.063	0.044	0.057	0.064	0.042
Total	0.041	0.043	0.030	0.040	0.044	0.029
RMSE						
Alive	0.056	0.059	0.053	0.057	0.059	0.043
Dead	0.061	0.063	0.049	0.057	0.064	0.043
Total	0.145	0.120	0.104	0.088	0.064	0.073

weighted sum of the component cell variances,[1] are larger than those from method (1) for larger categories but smaller for targeted hospitals and dead discharges.

[1]The cell variances are actually weighted by the squares of the population proportions. We use the shorthand reference to population weights because it keeps the sentence somewhere in the neighborhood of intelligibility.

Method (3) is unweighted estimates from the actual sample. The means for the four sample cells are still unbiased for this method, but the marginals are biased because the four sample cell means are combined using sample weights rather than population weights. The table shows the unweighted sample means for each cell and their biases (i.e., their difference from the population means in Table B.5). The bias will be small as long as the difference between the cell means is small or the difference between the sample and population proportions is small. This is true for addition across targeting categories to find means for dead or alive discharges. The unweighted sample means for targeted or untargeted hospitals, however, are substantially biased because dead discharges had substantially lower process scores and they were substantially overrepresented in the sample.

In principle, the bias in the unweighted estimates could be offset by lower standard errors. In fact, the standard errors are lower than those for the weighted estimates (they are equal to the inverse of the square root of the (actual) sample size), but not by enough to offset the effect of the bias. The combined effect of bias and standard error is measured by the root mean squared error (RMSE, calculated as the square root of the sum of the squares of the bias and the standard error). RMSE is larger for all of the marginal categories for method (3) than it is for method (2). (For method (2), RMSE is the same as standard error, because bias is 0.)

The unweighted estimates may yet have value, though, because the statistics in Table B.6 are not those we are ultimately interested in. What we want to know is the *difference* between targeted and untargeted, and between dead and alive. Table B.7 shows how the three methods stack up for measuring these differences.

The biases in the differences for method (3) are relatively small, because the biases in the two component categories are roughly the same. Equal biases in estimating the components subtract out and do not bias the estimate of the difference. And it is not just a fluke that the component biases are roughly the same. Method (3) underestimates average process of care in targeted and untargeted hospitals for exactly the same reasons: Dead patients are oversampled in both categories of hospitals, and dead patients have lower process scores in both categories of hospitals.

The standard errors for the differences (which are equal to the square root of the sum of squared standard errors for the two component categories) are smaller for the actual sample than for the proportional sample (substantially so for the dead/alive contrast). This is

Table B.7

COMPARISON OF THREE ESTIMATION METHODS: INPATIENT DEATHS AND TARGETING

	CHF Patients			AMI Patients		
	Method 1 Proportional	Method 2 Weighted	Method 3 Unweighted	Method 1 Proportional	Method 2 Weighted	Method 3 Unweighted
	Contrast: Targeted–Untargeted					
Bias	0.000	0.000	0.026	0.000	0.000	0.032
Standard error	0.074	0.072	0.060	0.094	0.065	0.059
RMSE	0.074	0.072	0.065	0.094	0.065	0.067
	Contrast: Dead–Alive					
Bias	0.000	0.000	−0.014	0.000	0.000	−0.014
Standard error	0.102	0.066	0.060	0.072	0.071	0.059
RMSE	0.102	0.066	0.061	0.072	0.071	0.061

because the standard error of the difference is more influenced by the component with the larger standard error. By equalizing standard errors across categories, our sampling plan minimizes the maximum standard error and so minimizes the standard error of the difference.

Combining bias and standard error as RMSE shows that in this case method (3) outperforms method (2), which in turn outperforms method (1), in estimating both the targeted/untargeted contrast and the dead/alive contrast. The differences in performance are fairly small with one exception: Our sample offers a large improvement over a proportional sample in estimating the difference in process of care received by dead and alive discharges.

Summary

Oversampling targeted hospitals and dead discharges was a success. Compared with a proportional sample, it substantially increased our ability to estimate differences between dead and alive discharges for CHF patients, and between patients in targeted and untargeted hospitals for AMI patients, without degrading our ability to estimate the other contrasts.

In the case investigated, this is true whether the estimates based on our sample are population weighted or not. The conclusion for weighted estimates should be quite robust and extend without much change to other situations. The weighted estimates are unbiased like

those from the proportional sample, so the comparison devolves to a comparison of standard errors. The reason that the conclusion is robust is that it depends mainly on the sample and population sizes in the four sampling cells, and these will be identical for other analyses of the data. It might be possible to reach different conclusions by postulating substantially greater standard deviations in the less populous sampling cells, but it seems unlikely that standard deviations actually differ enough across cells to reverse the conclusion.

The conclusion for the unweighted estimates is less robust, because it depends on the means in the sampling cells, not just the standard deviations. It should be fairly easy to construct a plausible example in which weighted estimates of the two differences have smaller RMSE than do unweighted estimates. But the weighted estimates are always feasible, so our sampling plan dominates proportional sampling for answering the two main study questions.

SAMPLE PERFORMANCE FOR ALTERNATIVE TARGETING

Background and Framework

The NOS was designed to test targeting that was completely nonobtrusive and easy to do. Targeting based on inpatient death can be done solely using claims data. Since the project started, though, targeting based on 30-day deaths has become prominent. This is harder to do because it requires merging death records with the claims data to track out-of-hospital deaths. But it is now being done routinely for the HCFA annual releases of hospital mortality data.

We would therefore like to be able to use our data to investigate the 30-day variants of the main study issues. In the context of this comparison of our sampling plan with a proportional sample, the two main questions are: (1) What is the difference between the average quality of care in hospitals that have higher than expected 30-day death rates and those that do not? (2) What is the difference between the average quality of care received by patients who die within 30 days of admission and those who live longer than that?

To investigate these questions, we merged social security death records with the claims records (using social security number and sex). We then used the merged file to identify hospitals with greater than expected 30-day deaths and patients who died within 30 days of admission. By this time, we were well into the field phase of the study, so it was too late to modify our sample to account for 30-day targeting.

Fortunately, unbiased estimates for the 30-day categories are easily obtained from our sample. The reason is that our stratified random sample is representative of the population within each of the four inpatient sampling categories. In particular, it is (ex ante) representative of the population distribution across the 30-day categories. For example, within the inpatient category consisting of patients discharged dead from targeted hospitals, the expected sample proportions equal population proportions for each of the four 30-day categories: targeted-dead, targeted-alive, untargeted-dead, untargeted-alive. Thus estimates using the same inpatient inverse sampling weights as used above will be ex ante unbiased.

Alternatively, we can apply separate population weights to each of the 16 cells obtained by crossing the four inpatient categories with the four 30-day categories. These weights account for sampling fluctuations that cause realized sample proportions to differ from population proportions, and result in estimates that are ex post unbiased, albeit at the cost of somewhat higher variance than inpatient weighted estimates. It is these 16-cell ex post weights whose performance we investigate here.

The Detailed Numbers

The results of the cross classification of inpatient and 30-day categories are shown in Tables B.8 and B.9. Here all four of the inpatient sampling categories are strung out as the four columns labeled td (for patients discharged dead from targeted hospitals), ta (targeted alive), ud (untargeted dead), and ua (untargeted alive). The corresponding 30-day categories are in rows with the same labels, but of course the row labels should be given their 30-day interpretations (for example, td is patients discharged from hospitals with more than expected deaths within 30 days of admission, who died within 30 days of admission). The resulting 16-cell matrix is bordered by a "total" row and "total" column, in the same manner as were the four-cell matrices in Tables B.5 and B.6. In addition, off to the right sides of Tables B.8 and B.9 are further aggregations into 30-day categories of interest (targeted and untargeted, dead and alive, all based on 30-day deaths).

There is a strong tendency for inpatient deaths and 30-day deaths to happen to the same people, and a weaker tendency for inpatient targeted hospitals to be also targeted for 30-day deaths. (Both tendencies are more easily seen in 2x2 tables not provided here. They are easy to construct from Tables B.8 and B.9. (See also Chassin et al., 1989.) The six most populous cells are the four on the diagonal, where the

Table B.8

POPULATION AND SAMPLES FOR SAMPLING COMPARISON: CHF 30-DAY DEATHS AND TARGETING[a]

	Inpatient td	Inpatient ta	Inpatient ud	Inpatient ua	Inpatient Total	Aggregate Categories[b]	
Population							
30-day td	777	88	210	211	1286	Targ30	7322
30-day ta	69	4466	0	1501	6036	Untarg30	74842
30-day ud	1361	256	4818	2986	9421	Dead30	10707
30-day ua	326	9119	192	55784	65421	Alive30	71457
30-day total	2533	13929	5220	60482	82164		
Actual sample							
30-day td	76	2	9	2	89	Targ30	198
30-day ta	9	92	0	8	109	Untarg30	928
30-day ud	134	5	248	15	402	Dead30	491
30-day ua	34	191	8	293	526	Alive30	635
30-day total	253	290	265	318	1126		
Raw weights							
30-day td	10.2	44.0	23.3	105.5	14.4	Targ30	37.0
30-day ta	7.7	48.5	0.0	187.6	55.4	Untarg30	80.6
30-day ud	10.2	51.2	19.4	199.1	23.4	Dead30	21.8
30-day ua	9.6	47.7	24.0	190.4	124.4	Alive30	112.5
30-day total	10.0	48.0	19.7	190.2	73.0		
Proportional sample							
30-day td	11	1	3	3	18	Targ30	100
30-day ta	1	61	0	21	83	Untarg30	1026
30-day ud	19	4	66	41	129	Dead30	147
30-day ua	4	125	3	764	897	Alive30	979
30-day total	35	191	72	829	1126		
Population means							
30-day td	−0.270	0.470	−0.700	−1.060	−0.419	Targ30	0.111
30-day ta	−0.040	0.210	0.000	0.280	0.225	Untarg30	0.090
30-day ud	−0.060	0.710	−0.260	−0.020	−0.129	Dead30	−0.164
30-day ua	−0.160	0.210	−0.110	0.110	0.122	Alive30	0.131
30-day total	−0.137	0.221	−0.272	0.104	0.092		

[a]Standard deviation is assumed equal to 1.00 in all cells.

[b]Targ30 denotes all patients in 30-day targeted hospitals; untarg30 denotes all patients in 30-day untargeted hospitals; dead30 denotes all patients who died within 30 days of admission; and alive30 denotes all patients still alive 30 days after admission.

inpatient and 30-day categorizations are the same, plus two others where about two-thirds of inpatient targeted dead and inpatient targeted alive become 30-day untargeted dead and alive, respectively.

Otherwise Tables B.8 and B.9 parallel Table B.5. The population means postulated for the 16 sampling categories are again sample means from our data, which are again unbiased estimates of the population means.

We compare the same three estimation methods as before.

1. Unweighted estimates based on the hypothetical proportional sample.
2. Weighted estimates based on our actual sample, using population weights for each of the 16 sampling cells.
3. Unweighted estimates using our actual sample.

Tables B.10 and B.11 show the details of the results for the three estimation methods. For method (1), unweighted sample means in each of the cells (including the "total" row and column) are unbiased estimates of the population means. The standard errors are just the inverse of the square root of the (proportional) sample sizes. The proportional sample provides the minimum variance estimate of the overall population mean, but the estimates for the less populous cells are less reliable. In particular, cells for 30-day targeted hospitals and dead within 30 days of admission have relatively high standard errors, and that is a clear disadvantage for the comparisons that we are making here.

Method (2), weighted estimates from the actual sample using 16 separate population weights, one for each of the sampling cells, also provides unbiased estimates. The marginal standard errors for the 30-day category row totals, calculated as the square root of the population weighted sum of the component cell variances, tend to be inflated by the population weighting. But that inflation tends to be offset by the oversampling of less populous cells, and for the smallest category (30-day targeted dead), the weighted estimates have a smaller standard error than does the proportional sample.

Method (3) is unweighted estimates from the actual sample. The means for the 16 sample cells are still unbiased for this method, but the marginals are biased because the 16 sample cell means are combined using sample weights rather than population weights. The table shows the unweighted sample means for each cell and their biases (i.e., their difference from the population means in Tables B.8 and B.9). The bias will be small as long as the difference between the cell means is small or the difference between the sample and population

Table B.9

POPULATION AND SAMPLES FOR SAMPLING COMPARISON: AMI 30-DAY DEATHS AND TARGETING[a]

	Inpatient td	Inpatient ta	Inpatient ud	Inpatient ua	Inpatient Total	Aggregate Categories[b]	
Population							
30-day td	753	85	182	0	1020	Targ30	2994
30-day ta	6	1608	0	360	1974	Untarg30	44926
30-day ud	799	54	8169	1393	10415	Dead30	11435
30-day ua	32	1920	188	32371	34511	Alive30	36485
30-day total	1590	3667	8539	34124	47920		
Actual sample							
30-day td	116	4	8	0	128	Targ30	257
30-day ta	1	127	0	1	129	Untarg30	893
30-day ud	121	2	297	9	429	Dead30	557
30-day ua	5	152	6	301	464	Alive30	593
30-day total	243	285	311	311	1150		
Raw weights							
30-day td	6.5	21.3	22.8	0.0	8.0	Targ30	11.6
30-day ta	6.0	12.7	0.0	360.0	15.3	Untarg30	50.3
30-day ud	6.6	27.0	27.5	154.8	24.3	Dead30	20.5
30-day ua	6.4	12.6	31.3	107.5	74.4	Alive30	61.5
30-day total	6.5	12.9	27.5	109.7	41.7		
Proportional sample							
30-day td	18	2	4	0	24	Targ30	72
30-day ta	0	39	0	9	47	Untarg30	1078
30-day ud	19	1	196	33	250	Dead30	274
30-day ua	1	46	5	777	828	Alive30	876
30-day total	38	88	205	819	1150		
Population means							
30-day td	0.120	0.796	−0.019	0.000	0.152	Targ30	0.320
30-day ta	1.174	0.388	0.000	0.482	0.408	Untarg30	0.262
30-day ud	−0.040	1.358	0.057	0.530	0.120	Dead30	0.122
30-day ua	0.217	0.287	−0.110	0.308	0.304	Alive30	0.310
30-day total	0.046	0.359	0.052	0.319	0.265		

[a]Standard deviation is assumed equal to 1.00 in all cells.

[b]Targ30 denotes all patients in 30-day targeted hospitals; untarg30 denotes all patients in 30-day untargeted hospitals; dead30 denotes all patients who died within 30 days of admission; and alive30 denotes all patients still alive 30 days after admission.

Table B.10

ESTIMATES FOR SAMPLE COMPARISON: CHF 30-DAY DEATHS AND TARGETING

	Inpatient td	Inpatient ta	Inpatient ud	Inpatient ua	Inpatient Total	Aggregate Categories[a]	
(1) Unweighted Estimates from Proportional Sample Would Be Unbiased							
Standard errors							
30-day td	0.306	0.911	0.589	0.588	0.238	Targ30	0.100
30-day ta	1.028	0.128	—	0.220	0.110	Untarg30	0.031
30-day ud	0.232	0.534	0.123	0.156	0.088	Dead30	0.083
30-day ua	0.473	0.089	0.616	0.036	0.033	Alive30	0.032
30-day total	0.170	0.072	0.118	0.035	0.030		
(2) 16-Cell Weighted Estimates from Actual Sample Are Also Unbiased							
Standard errors							
30-day td	0.115	0.707	0.333	0.707	0.154	Targ30	0.100
30-day ta	0.333	0.104	0.000	0.354	0.117	Untarg30	0.046
30-day ud	0.086	0.447	0.064	0.258	0.090	Dead30	0.081
30-day ua	0.171	0.072	0.354	0.058	0.051	Alive30	0.048
30-day total	0.063	0.059	0.062	0.056	0.043		
(3) Unweighted Means from Actual Sample Are Biased but May Have Smaller RMSE							
Expected sample means							
30-day td	−0.270	0.470	−0.700	−1.060	−0.315	Targ30	−0.034
30-day ta	−0.040	0.210	0.000	0.280	0.194	Untarg30	−0.004
30-day ud	−0.060	0.710	−0.260	−0.020	−0.172	Dead30	−0.198
30-day ua	−0.160	0.210	−0.110	0.110	0.126	Alive30	0.137
30-day total	−0.136	0.220	−0.270	0.101	−0.009		
Bias							
30-day td	0.000	0.000	0.000	0.000	0.105	Targ30	−0.146
30-day ta	0.000	0.000	0.000	0.000	−0.030	Untarg30	−0.094
30-day ud	0.000	0.000	0.000	0.000	−0.044	Dead30	−0.035
30-day ua	0.000	0.000	0.000	0.000	0.004	Alive30	0.007
30-day total	0.001	−0.000	0.002	−0.003	−0.101		
Standard errors							
30-day td	0.115	0.707	0.333	0.707	0.106	Targ30	0.071
30-day ta	0.333	0.104	0.000	0.354	0.096	Untarg30	0.033
30-day ud	0.086	0.447	0.064	0.258	0.050	Dead30	0.045
30-day ua	0.171	0.072	0.354	0.058	0.044	Alive30	0.040
30-day total	0.063	0.059	0.061	0.056	0.030		
RMSE							
30-day td	0.115	0.707	0.333	0.707	0.149	Targ30	0.162
30-day ta	0.333	0.104	0.000	0.354	0.100	Untarg30	0.099
30-day ud	0.086	0.447	0.064	0.258	0.066	Dead30	0.057
30-day ua	0.171	0.072	0.354	0.058	0.044	Alive30	0.040
30-day total	0.063	0.059	0.061	0.056	0.106		

[a]Targ30 denotes all patients in 30-day targeted hospitals; untarg30 denotes all patients in 30-day untargeted hospitals; dead30 denotes all patients who died within 30 days of admission; and alive30 denotes all patients still alive 30 days after admission.

Table B.11

ESTIMATES FOR SAMPLE COMPARISON: AMI 30-DAY DEATHS AND TARGETING

	Inpatient td	Inpatient ta	Inpatient ud	Inpatient ua	Inpatient Total	Aggregate Categories[a]	
(1) Unweighted Estimates from Proportional Sample Would Be Unbiased							
Standard errors							
30-day td	0.235	0.700	0.478	0.000	0.202	Targ30	0.118
30-day ta	2.635	0.161	0.000	0.340	0.145	Untarg30	0.030
30-day ud	0.228	0.878	0.071	0.173	0.063	Dead30	0.060
30-day ua	1.141	0.147	0.471	0.036	0.035	Alive30	0.034
30-day total	0.162	0.107	0.070	0.035	0.029		
(2) 16-Cell Weighted Estimates from Actual Sample Are Also Unbiased							
Standard errors							
30-day td	0.093	0.500	0.354	0.000	0.102	Targ30	0.134
30-day ta	1.000	0.089	0.000	1.000	0.196	Untarg30	0.044
30-day ud	0.091	0.707	0.058	0.333	0.064	Dead30	0.059
30-day ua	0.447	0.081	0.408	0.058	0.054	Alive30	0.052
30-day total	0.064	0.060	0.057	0.057	0.042		
(3) Unweighted Means from Actual Sample Are Biased but May Have Smaller RMSE							
Expected sample means							
30-day td	0.120	0.796	−0.019	0.000	0.132	Targ30	0.264
30-day ta	1.174	0.388	0.000	0.482	0.395	Untarg30	0.175
30-day ud	−0.040	1.358	0.057	0.530	0.046	Dead30	0.066
30-day ua	0.217	0.287	−0.110	0.308	0.295	Alive30	0.317
30-day total	0.047	0.347	0.052	0.315	0.195		
Bias							
30-day td	0.000	0.000	0.000	0.000	−0.019	Targ30	−0.056
30-day ta	0.000	0.000	0.000	0.000	−0.013	Untarg30	−0.087
30-day ud	0.000	0.000	0.000	0.000	−0.074	Dead30	−0.057
30-day ua	0.000	0.000	0.000	0.000	−0.010	Alive30	0.006
30-day total	0.001	−0.012	0.000	−0.004	−0.070		
Sandard errors							
30-day td	0.093	0.500	0.354	0.000	0.088	Targ30	0.062
30-day ta	1.000	0.089	0.000	1.000	0.088	Untarg30	0.033
30-day ud	0.091	0.707	0.058	0.333	0.048	Dead30	0.042
30-day ua	0.447	0.081	0.408	0.058	0.046	Alive30	0.041
30-day total	0.064	0.059	0.057	0.057	0.029		
RMSE							
30-day td	0.093	0.500	0.354	0.000	0.090	Targ30	0.084
30-day ta	1.000	0.089	0.000	1.000	0.089	Untarg30	0.093
30-day ud	0.091	0.707	0.058	0.333	0.088	Dead30	0.071
30-day ua	0.447	0.081	0.408	0.058	0.047	Alive30	0.042
30-day total	0.064	0.060	0.057	0.057	0.076		

[a]Targ30 denotes all patients in 30-day targeted hospitals; untarg30 denotes all patients in 30-day untargeted hospitals; dead30 denotes all patients who died within 30 days of admission; and alive30 denotes all patients still alive 30 days after admission.

proportions is small. The extremely small biases (ex post sampling) shown in the four inpatient targeting columns arise solely because of sampling variation in the proportion of hospitalizations in each 30-day targeting category; the expected bias for these four columns is zero. The bias in the 30-day targeting row totals is only moderate.

In principle, the bias in the unweighted estimates could be offset by lower standard errors. In fact, in this case it is more than offset. The combined effect of bias and standard error as measured by RMSE is smaller for all of the marginal categories for method (3) than it is for method (2).

Table B.12 shows how the three methods compare in estimating the contrasts between 30-day targeted and untargeted hospitals, and between 30-day dead and alive patients. Somewhat surprisingly, the unbiased method (2) estimates using 16-cell population weights have only slightly higher standard errors than would the method (1) estimates based on a hypothetical proportional sample. Because that conclusion rests largely on the relative sample and population sizes in the 16 cells, it is fairly robust. In particular, it does not depend in any way on the values of the 16-cell means or the way in which they are distributed across the cells.

In this case, the unweighted estimates (method (3)) have lower RMSEs than either methods (1) or (2) but the differences are again quite small. This conclusion does depend on the values of the sample means; the biases might be much larger for some other sets of sample means.

OVERALL SUMMARY

Judged by its ability to answer our two original study questions (comparisons of care received (1) by patients in inpatient targeted and untargeted hospitals and (2) by patients who died in the hospital and those who lived), our sample is a clear winner over proportional sampling. For the comparison of CHF patients discharged dead or alive, a weighted (unbiased) estimate in our sample reduces standard error by 35 percent relative to a proportional sample. For the comparison of AMI patients in targeted or untargeted hospitals, it reduces standard error by 30 percent. And it achieves these improvements without decreasing precision for the other two central comparisons (targeted/untargeted for CHF, dead/alive for AMI).

The increase in efficiency for inpatient estimates is bought at the cost of a smaller decrease in efficiency for 30-day estimates. Again focusing on weighted (unbiased) estimates, the increase in standard

Table B.12

COMPARISON OF THREE ESTIMATION METHODS: INPATIENT AND 30-DAY DEATHS AND TARGETING

	CHF Patients			AMI Patients		
	Method 1 Proportional	Method 2 Weighted	Method 3 Unweighted	Method 1 Proportional	Method 2 Weighted	Method 3 Unweighted
Inpatient Deaths and Targeting						
Contrast: Targeted–Untargeted						
Bias	0.000	0.000	0.026	0.000	0.000	0.032
Standard error	0.074	0.072	0.060	0.094	0.065	0.059
RMSE	0.074	0.072	0.065	0.094	0.065	0.067
Contrast: Dead–Alive						
Bias	0.000	0.000	−0.014	0.000	0.000	−0.014
Standard error	0.102	0.066	0.060	0.072	0.071	0.059
RMSE	0.102	0.066	0.061	0.072	0.071	0.061
30-Day Deaths and Targeting						
Contrast: Targeted–Untargeted						
Bias	0.000	0.000	−0.052	0.000	0.000	0.030
Standard error	0.105	0.110	0.078	0.122	0.141	0.071
RMSE	0.105	0.110	0.094	0.122	0.141	0.077
Contrast: Dead–Alive						
Bias	0.000	0.000	−0.041	0.000	0.000	−0.063
Standard error	0.089	0.094	0.060	0.069	0.079	0.059
RMSE	0.089	0.094	0.073	0.069	0.079	0.087

NOTE: Results for inpatient deaths are brought forward from Table B.7 for convenient reference.

error in our sample relative to a proportional sample ranges from about 5 percent for the two CHF contrasts to about 15 percent for the two AMI contrasts.

Our sample will be better for *all* contrasts of interest in situations where unweighted estimates are appropriate, i.e., when the differences in sample cell values are not too large. The unweighted estimates, though biased, may have smaller RMSEs than the unbiased weighted estimates. In our analysis, we calculate both weighted and unweighted estimates and rely more on the (more efficient) unweighted estimates when the two do not differ substantially (see Hausman, 1978).

Appendix C

BINOMIAL SIMULATION OF OUTCOME TARGETING IN FOUR SAMPLE STATES

Table C.1 shows that death rates in targeted hospitals are substantially higher than those in untargeted hospitals, ranging from 40 percent higher for CHF 30-day deaths to almost 100 percent higher for CHF inpatient deaths. For AMI patients, targeted hospitals have about 50 percent higher death rates, regardless of whether deaths are counted in the hospital or within 30 days of admission.

Table C.1

SIMULATION RESULTS FOR 1137 HOSPITALS TREATING CHF PATIENTS AND 1121 HOSPITALS TREATING AMI PATIENTS IN FOUR SAMPLE STATES

	Untargeted Hospitals	Targeted Hospitals	Difference	Chi-Squared
		CHF Patients		
Inpatient death rates				
Actual	7.9	15.4	7.4	2783
Simulated	7.8	12.9	5.1	1040
	(0.01)	(0.02)	(0.02)	(6)
30-day death rates				
Actual	12.6	17.6	5.0	1711
Simulated	12.7	16.8	4.1	1140
	(0.01)	(0.03)	(0.03)	(5)
		AMI Patients		
Inpatient death rates				
Actual	20.0	30.2	10.2	2193
Simulated	20.3	26.7	6.3	1120
	(0.02)	(0.05)	(0.05)	(5)
30-day death rates				
Actual	23.2	34.1	10.9	1977
Simulated	24.5	30.6	6.1	1119
	(0.02)	(0.07)	(0.06)	(5)

NOTE: Simulated values are means from 100 trials; standard errors are in parentheses.

The higher death rates in targeted hospitals are to some extent an inevitable result of the way in which the hospitals are targeted; they are targeted precisely because they have higher than expected death rates. The simulation results show how much of the difference in death rates can be attributed solely to the targeting method. Even if hospitals differed only in the age/sex/race mix of their patients, targeted hospitals would have death rates that ranged from 25 to 65 percent higher than untargeted hospitals.

We simulated the design effect as follows (for concreteness, the description uses CHF inpatient deaths as an example; the simulations for 30-day deaths and for AMI patients used the same method). For all of the hospitals in the four NOS sample states, we simulated their CHF inpatient deaths on the null hypothesis that the probability of dying in each hospital, p_i, was what it would have been if national average death rates for each age/sex/race cell applied to that hospital's age/sex/race distribution of patients. For each patient discharged from each hospital, we generated a random number from a uniform distribution over the interval zero to one. If that number was smaller than p_i for that hospital, we counted that patient as discharged dead in the simulation; if it was greater, we counted the patient as alive. We added up the simulated deaths for each hospital, getting a number that might be zero or might be some positive integer. We then ranked the hospitals by the binomial probability that they would have as many as the simulated number of deaths, and designated the 145 hospitals (out of 1137 total) with the lowest probabilities as simulated targeted hospitals. (145 is the number of hospitals actually targeted.) We calculated the spread between the death rates in simulated targeted hospitals as a group and simulated untargeted hospitals as a group. We repeated the process 100 times. The mean spread was (for CHF inpatient deaths) 5.1 percentage points with a standard error of the mean of only 0.02.

Because most hospitals had too few expected deaths to meet the usual statistical criteria for employing a chi-squared test, we also used the simulation model to generate distributions of the chi-squared statistic for the mortality data represented in each condition. The reasoning for this strategy is as follows. The sum of squares of N independent unit normal random deviates is distributed as chi-squared with N degrees of freedom. The commonly used normal approximation to the binomial distribution would, in our case, treat the number of deaths m_i for a given hospital i as a normally distributed random variable with mean $n_i p_i$ and variance $n_i p_i(1 - p_i)$. This suggests that we could use as a test statistic the following sum of squares of approximately unit normally distributed quantities:

$$\sum \frac{(m_i - n_i p_i)^2}{(n_i p_i (1 - p_i))}.$$

However, the normal approximation is not very good when the number of expected events, $E(m_i) = n_i p_i$, is small. A common rule of thumb warns against relying on a chi-squared test when more than 20 percent of the cells contain fewer than five expected events, as was the case for both CHF and AMI. Thus, it seemed unwise to rely on a standard chi-squared test. Instead, we used the computer simulation to estimate the true distribution of the test statistic described above, on the null hypothesis that the probability of death in each hospital equaled its expected death rate. Substituting simulated deaths for each hospital for m_i, we calculated a value for the test statistic.

Random variation and the selection of targeted hospitals account for a large part of the differences in death rates, but not all of the differences, and the nonrandom parts are highly significant statistically. The empirical distributions of the test statistic, on the simulation null hypothesis, are summarized by the means and standard deviations shown in the table. In all cases, the actual value of the statistic is at least 10 standard deviations above the mean, indicating that random variation alone cannot account for observed differences in death rates between targeted and untargeted hospitals.

Here are some additional untabulated results from the simulation of CHF inpatient deaths: Only 39.6 hospitals on average (standard deviation 5.6, range 27 to 53) would have been targeted if we had used a 0.05 probability cut for the simulated targeting instead of a 145-lowest-probability cut. The mean probability for the highest-probability simulated targeted hospital (i.e., the last hospital to make the cut) was 0.170 (standard deviation 0.012, range 0.130 to 0.196).

39.6 out of 1137 hospitals constitutes only 3.5 percent, significantly less than the 5 percent that one might expect when using a 0.05 probability cutoff. This is because of the granularity of the binomial distribution. For example, presume a hospital with $n = 5$ CHF patients. Assuming that each has a probability of dying of 0.1, the probability that two or more of them die is 0.0815, not sufficiently low to get it targeted. But the probability of three or more dying is only 0.0086, so on the null hypothesis it would be targeted at the 0.05 level less than 1 percent of the time. The granularity smooths out as n increases, but surprisingly slowly, as shown in Fig. C.1. (The figure shows a plot of p against n, where n is the number of patients and p is the probability on the null hypothesis that a hospital with that many patients would be targeted using a 0.05 probability cutoff. For this plot, the probability of death is assumed to be 0.1 in all hospitals.)

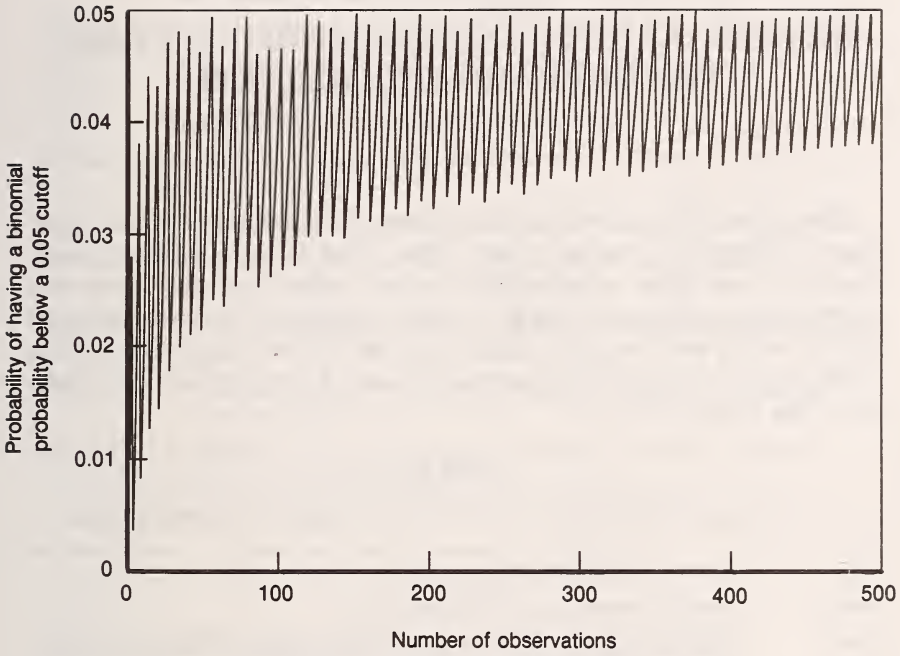


Fig. C.1—Illustrating granularity of the binomial distribution

The average actually targeted hospital is much larger (160 CHF patients, standard deviation 100, range 3 to 487) than is the average actually untargeted hospital (mean 85, standard deviation 71, range 1 to 532). This is mainly a real rather than a design effect: Simulated targeted hospitals averaged 103 CHF patients; simulated untargeted hospitals averaged 84. One would expect some design effect because the power to target larger hospitals is greater than the power to target smaller hospitals.

Appendix D

SEVERE ILLNESS AND BAD CARE INCREASE THE PROBABILITY OF DEATH

This appendix consists largely of alternative estimates of the equations in Table 4 in the main text. The point is that alternative estimates do not differ substantially, except where the differences are appropriately reflected in Table 4. Also included are estimates of state effects (which are frequently significant; see Table D.10) and estimates of the effects of hospital characteristics (only a few of which are significant; see Table D.11).

Table D.1

VARIABLE DEFINITIONS FOR ESTIMATING THE RECURSIVE MODEL

Age	Patient's age in decades
Severity	Patient's severity score (PPS scale/100)
DNR	Indicator variable equals one if DNR order written for the patient on the first day in the hospital
Quality	Patient's summary quality of process score (PPS scale)
log(LOS)	Natural log of patient's length of stay (days in the hospital)
Home	Indicator variable equals one in the Cox model for days following hospital discharge and 30 days or less after admission
Deadin	Patient died in hospital
Dead30	Patient died within 30 days of admission
Constant	Constant term
R-square	Unadjusted R-squared
Observations	Number of observations

NOTE: Both severity and DNR are scaled smaller by a factor of 100 in the regressions (Table 4 and Appendix D) than they are in the comparison tables (Tables 5 and 6, and Appendix E).

Table D.2

KEY TO ALTERNATIVE REGRESSION ESTIMATES

Number	Description
(1) Unwt	Unweighted regression on essential variables only
(2) Wtin	Weighted regression on essential variables only
(3) States	Unweighted regression on essential variables plus indicator variables for sample states
(4) Hchar	Unweighted regression on essential variables plus hospital characteristics
(5) Alive	Unweighted regression on essential variables for patients discharged from the hospital alive
(6) Deadin	Unweighted regression on essential variables for patients who died in the hospital
(7) Untarg	Unweighted regression on essential variables for patients in inpatient targeted hospitals
(8) Targin	Unweighted regression on essential variables for patients in hospitals not targeted for inpatient deaths

Table D.3

ALTERNATIVE LOGISTIC ESTIMATES OF DNR STATUS ON THE FIRST DAY OF ADMISSION

	(1) Unwt	(2) Wtin	(3) States	(4) Hchar	(5) Alive	(6) Deadin	(7) Untarg	(8) Targin
				CHF Patients				
Age	0.36	0.46	0.45	0.38	0.34	0.35	0.49	0.12
	(2.3)	(2.1)	(2.8)	(2.4)	(0.8)	(2.0)	(2.8)	(0.4)
Severity	8.43	11.44	8.05	8.38	14.02	4.61	7.74	9.82
	(6.3)	(5.9)	(5.9)	(6.2)	(3.4)	(3.0)	(5.0)	(3.5)
Constant	−8.94	−11.27	−8.86	−9.07	−11.89	−6.76	−9.21	−8.53
	(−7.0)	(−6.2)	(−6.7)	(−6.8)	(−3.2)	(−4.8)	(−6.1)	(−3.3)
R-square	0.05	0.05	0.09	0.08	0.03	0.03	0.07	0.01
Observations	1126	1126	1126	1125	608	518	583	543
				AMI Patients				
Age	0.79	1.04	0.76	0.73	1.97	0.69	0.76	0.82
	(3.9)	(3.8)	(3.7)	(3.5)	(1.8)	(3.3)	(3.0)	(2.4)
Severity	4.88	6.26	4.90	5.46	9.73	3.09	5.12	4.61
	(6.1)	(5.7)	(5.9)	(6.2)	(1.9)	(3.5)	(4.8)	(3.6)
Constant	−11.39	−14.21	−10.64	−10.80	−24.49	−9.38	−11.10	−11.72
	(−6.8)	(−6.2)	(−6.2)	(−6.1)	(−2.6)	(−5.4)	(−5.3)	(−4.2)
R-square	0.06	0.08	0.09	0.12	0.04	0.05	0.10	0.02
Observations	1149	1149	1149	1149	596	553	621	528

NOTE: See Table D.2 for definitions of the column headings.

Table D.4

ALTERNATIVE ORDINARY LEAST SQUARES REGRESSION ESTIMATES OF QUALITY OF PROCESS OF CARE

	(1) Unwt	(2) Wtin	(3) States	(4) Hchar	(5) Alive	(6) Deadin	(7) Untarg	(8) Targin
				CHF Patients				
Age	−0.12	−0.09	−0.13	−0.12	−0.12	−0.13	−0.10	−0.15
	(−3.8)	(−3.1)	(−4.0)	(−3.8)	(−2.8)	(−2.5)	(−2.2)	(−3.3)
Severity	−1.06	−0.50	−1.06	−1.07	−0.07	−0.76	−1.06	−0.98
	(−3.3)	(−1.5)	(−3.4)	(−3.4)	(−0.1)	(−1.5)	(−2.5)	(−2.1)
DNR	−0.55	−0.52	−0.50	−0.49	−0.39	−0.51	−0.50	−0.64
	(−4.5)	(−3.3)	(−4.1)	(−4.1)	(−1.3)	(−3.5)	(−3.5)	(−2.4)
Constant	1.32	0.96	1.13	1.37	1.04	1.13	1.10	1.58
	(5.4)	(4.2)	(4.7)	(5.6)	(3.4)	(2.7)	(3.2)	(4.6)
R-square	0.06	0.03	0.10	0.10	0.02	0.05	0.06	0.05
Observations	1126	1126	1126	1125	608	518	583	543
				AMI Patients				
Age	−0.18	−0.16	−0.17	−0.19	−0.18	−0.21	−0.22	−0.14
	(−5.3)	(−4.7)	(−5.2)	(−5.5)	(−4.1)	(−3.9)	(−4.7)	(−2.7)
Severity	−0.83	−0.53	−0.89	−0.77	0.15	−1.15	−0.80	−0.88
	(−5.5)	(−3.2)	(−6.0)	(−5.1)	(0.5)	(−5.2)	(−3.8)	(−4.0)
DNR	−0.92	−1.19	−0.89	−0.91	−1.47	−0.83	−1.09	−0.62
	(−7.1)	(−7.7)	(−7.0)	(−7.1)	(−3.4)	(−5.6)	(−6.7)	(−2.9)
Constant	1.88	1.62	1.66	1.92	1.71	2.17	2.17	1.54
	(7.3)	(6.5)	(6.4)	(7.3)	(5.2)	(5.4)	(6.2)	(4.1)
R-square	0.13	0.11	0.17	0.16	0.05	0.16	0.17	0.09
Observations	1149	1149	1149	1149	596	553	621	528

NOTE: See Table D.2 for definitions of the column headings.

Table D.5

ALTERNATIVE ORDINARY LEAST SQUARES REGRESSION ESTIMATES OF THE LOGARITHM OF LENGTH OF STAY

	(1) Unwt	(2) Wtin	(3) States	(4) Hchar	(5) Alive	(6) Deadin	(7) Untarg	(8) Targin
				CHF Patients				
Age	0.05	0.07	0.04	0.05	0.04	0.05	0.08	−0.01
	(1.8)	(3.0)	(1.2)	(1.8)	(1.4)	(1.0)	(2.2)	(−0.3)
Severity	−0.79	0.13	−0.62	−0.77	1.00	−2.23	−1.21	0.00
	(−2.7)	(0.5)	(−2.2)	(−2.6)	(2.7)	(−4.3)	(−3.4)	(0.0)
DNR	−0.61	−0.28	−0.47	−0.59	0.04	−0.65	−0.36	−1.04
	(−5.4)	(−2.2)	(−4.2)	(−5.2)	(0.2)	(−4.4)	(−3.0)	(−4.1)
Process	0.08	−0.01	0.08	0.06	−0.01	0.15	0.09	0.05
	(2.9)	(−0.5)	(3.0)	(2.2)	(−0.3)	(3.3)	(2.6)	(1.1)
Constant	2.04	1.55	2.08	2.03	1.57	2.68	1.79	2.46
	(9.0)	(8.7)	(9.4)	(8.8)	(6.8)	(6.3)	(6.3)	(7.2)
R-square	0.05	0.01	0.11	0.08	0.02	0.11	0.06	0.04
Observations	1126	1126	1126	1125	608	518	583	543
				AMI Patients				
Age	0.08	0.11	0.08	0.07	0.02	0.07	0.07	0.09
	(2.2)	(3.3)	(2.4)	(2.0)	(0.8)	(1.5)	(1.5)	(1.6)
Severity	−2.80	−2.00	−2.73	−2.75	0.43	−2.14	−2.57	−3.05
	(−18.1)	(−12.7)	(−17.8)	(−17.7)	(2.2)	(−10.0)	(−12.2)	(−13.4)
DNR	−0.37	−0.59	−0.32	−0.38	−0.14	−0.11	−0.54	−0.12
	(−2.7)	(−4.0)	(−2.4)	(−2.8)	(−0.5)	(−0.8)	(−3.2)	(−0.5)
Process	0.16	0.11	0.17	0.14	0.06	0.18	0.12	0.19
	(5.3)	(3.9)	(5.6)	(4.7)	(2.1)	(4.3)	(3.1)	(4.3)
Constant	2.14	1.91	2.03	2.15	2.25	1.52	2.09	2.20
	(8.1)	(8.1)	(7.6)	(7.9)	(10.0)	(3.9)	(5.8)	(5.6)
R-square	0.30	0.18	0.31	0.31	0.02	0.23	0.28	0.32
Observations	1149	1149	1149	1149	596	553	621	528

NOTE: See Table D.2 for definitions of the column headings.

Table D.6

ALTERNATIVE LOGISTIC REGRESSION ESTIMATES OF INPATIENT DEATH

	(1) Unwt	(2) Wtin	(3) States	(4) Hchar	(7) Untarg	(8) Targin
			CHF Patients			
Age	0.05	0.09	0.03	0.04	0.01	0.07
	(0.6)	(0.7)	(0.4)	(0.5)	(0.1)	(0.5)
Severity	13.32	11.18	13.62	13.48	12.12	15.36
	(13.3)	(8.6)	(13.4)	(13.3)	(9.3)	(9.7)
DNR	1.31	1.03	1.49	1.32	1.46	1.43
	(3.2)	(2.5)	(3.6)	(3.2)	(3.3)	(1.3)
Process	−0.29	−0.27	−0.29	−0.29	−0.26	−0.34
	(−3.8)	(−2.2)	(−3.7)	(−3.7)	(−2.5)	(−3.1)
Constant	−5.35	−7.03	−5.39	−5.21	−4.81	−5.99
	(−8.0)	(−6.7)	(−8.0)	(−7.6)	(−5.3)	(−6.0)
R-square	0.27	0.13	0.28	0.28	0.28	0.27
Observations	1126	1126	1126	1125	583	543
			AMI Patients			
Age	−0.01	0.04	−0.01	−0.02	0.14	−0.21
	(−0.1)	(0.3)	(−0.1)	(−0.2)	(1.0)	(−1.4)
Severity	7.79	7.65	7.85	7.79	7.40	8.32
	(13.5)	(12.0)	(13.6)	(13.5)	(9.6)	(9.5)
DNR	1.75	1.04	1.78	1.73	1.25	—
	(2.7)	(2.1)	(2.8)	(2.7)	(1.8)	
Process	−0.06	−0.05	−0.05	−0.07	0.02	−0.17
	(−0.7)	(−0.5)	(−0.5)	(−0.8)	(0.2)	(−1.4)
Constant	−2.22	−3.84	−2.16	−2.01	−3.26	−0.94
	(−3.0)	(−4.3)	(−2.8)	(−2.6)	(−3.2)	(−0.8)
R-square	0.27	0.24	0.27	0.27	0.25	0.28
Observations	1149	1149	1149	1149	621	511

NOTE: See Table D.2 for definitions of the column headings.

Table D.7

ALTERNATIVE LOGISTIC REGRESSION ESTIMATES OF DEATH WITHIN 30 DAYS OF ADMISSION

	(1) Unwt	(2) Wt30	(3) States	(4) Hchar	(7) Untarg	(8) Targ30
			CHF Patients			
Age	−0.02	−0.15	−0.02	−0.02	−0.04	0.20
	(−0.2)	(−1.3)	(−0.2)	(−0.2)	(−0.4)	(0.9)
Severity	12.99	11.89	13.06	13.13	12.54	15.60
	(13.2)	(9.7)	(13.2)	(13.2)	(11.8)	(5.9)
DNR	1.85	2.15	1.88	1.85	1.75	—
	(4.1)	(4.8)	(4.1)	(4.1)	(3.8)	
Process	−0.24	−0.23	−0.23	−0.22	−0.19	−0.49
	(−3.1)	(−2.1)	(−3.0)	(−2.8)	(−2.2)	(−2.7)
Constant	−4.87	−5.08	−4.80	−4.87	−4.53	−7.50
	(−7.4)	(−5.5)	(−7.2)	(−7.1)	(−6.3)	(−4.0)
R-square	0.27	0.20	0.27	0.27	0.26	0.32
Observations	1126	1126	1126	1125	928	190
			AMI Patients			
Age	0.03	0.14	0.03	0.03	−0.01	0.17
	(0.3)	(1.3)	(0.3)	(0.3)	(−0.1)	(0.8)
Severity	7.49	6.53	7.52	7.48	7.52	7.45
	(13.3)	(11.1)	(13.3)	(13.3)	(11.8)	(6.2)
DNR	1.45	1.02	1.46	1.46	1.17	—
	(2.5)	(2.1)	(2.5)	(2.5)	(2.0)	
Process	−0.01	0.05	0.00	−0.01	−0.02	0.01
	(−0.2)	(0.5)	(0.0)	(−0.1)	(−0.2)	(0.0)
Constant	−2.45	−4.12	−2.38	−2.38	−2.18	−3.54
	(−3.3)	(−4.9)	(−3.1)	(−3.1)	(−2.6)	(−2.1)
R-square	0.25	0.19	0.25	0.25	0.25	0.25
Observations	1149	1149	1149	1149	892	247

NOTE: See Table D.2 for definitions of the column headings.

Table D.8

ALTERNATIVE COX PROPORTIONAL HAZARD ESTIMATES OF INPATIENT DEATH

	(1) Unwt	(3) States	(4) Hchar	(7) Untarg	(8) Targin
			CHF Patients		
Age	0.03	0.07	0.03	0.02	0.09
	(0.6)	(1.3)	(0.6)	(0.3)	(1.2)
Severity	5.42	5.48	5.68	6.31	4.62
	(11.8)	(11.8)	(12.1)	(9.6)	(7.1)
DNR	0.86	0.65	0.82	0.58	1.26
	(5.8)	(4.3)	(5.5)	(3.3)	(4.0)
Process	−0.12	−0.11	−0.11	−0.14	−0.09
	(−2.6)	(−2.3)	(−2.3)	(−2.0)	(−1.4)
Observations	1126	1126	1125	583	543
			AMI Patients		
Age	−0.01	−0.02	−0.01	0.02	−0.08
	(−0.2)	(−0.3)	(−0.1)	(0.2)	(−0.9)
Severity	4.58	4.57	4.58	4.19	5.07
	(18.2)	(17.9)	(17.9)	(12.5)	(13.0)
DNR	0.35	0.33	0.34	0.58	0.18
	(2.0)	(1.9)	(2.0)	(2.5)	(0.7)
Process	−0.16	−0.15	−0.15	−0.11	−0.18
	(−3.4)	(−3.2)	(−3.1)	(−1.8)	(−2.4)
Observations	1149	1149	1149	621	528

NOTE: See Table D.2 for definitions of the column headings.

Table D.9

ALTERNATIVE COX PROPORTIONAL HAZARD ESTIMATES OF DEATH WITHIN 30 DAYS OF ADMISSION

	(1) Unwt	(3) States	(4) Hchar	(7) Untarg	(8) Targ30
			CHF Patients		
Age	−0.02	0.01	−0.01	−0.04	0.03
	(−0.3)	(0.2)	(−0.2)	(−0.6)	(0.2)
Severity	5.86	5.89	6.12	5.63	7.76
	(12.5)	(12.5)	(12.8)	(11.1)	(6.0)
DNR	0.94	0.78	0.90	1.02	0.45
	(6.4)	(5.1)	(6.1)	(6.4)	(1.1)
Process	−0.11	−0.10	−0.09	−0.08	−0.27
	(−2.2)	(−1.9)	(−1.9)	(−1.5)	(−2.4)
Home	−2.62	−2.77	−2.68	−2.64	−2.51
	(−11.7)	(−12.3)	(−12.0)	(−10.8)	(−4.6)
Observations	1714	1714	1712	1415	299
			AMI Patients		
Age	−0.02	−0.02	−0.01	−0.03	0.05
	(−0.3)	(−0.3)	(−0.2)	(−0.5)	(0.4)
Severity	4.60	4.59	4.60	4.58	4.72
	(18.3)	(18.1)	(18.0)	(16.2)	(8.5)
DNR	0.42	0.40	0.42	0.44	0.35
	(2.5)	(2.3)	(2.4)	(2.2)	(1.0)
Process	−0.13	−0.13	−0.12	−0.14	−0.13
	(−2.9)	(−2.7)	(−2.6)	(−2.6)	(−1.2)
Home	−2.26	−2.30	−2.30	−2.40	−1.82
	(−7.4)	(−7.4)	(−7.5)	(−6.7)	(−2.9)
Observations	1727	1727	1727	1345	382

NOTE: See Table D.2 for definitions of the column headings.

Table D.10

REGRESSION RESULTS FOR STATES

State	Logit DNR	OLS Quality	OLS log(LOS) Alive	OLS log(LOS) Deadin	Logit Deadin	Logit Dead30	Cox Deadin	Cox Dead30
				CHF Patients				
California	−0.12	0.46	−0.38	−0.10	−0.19	−0.21	0.05	0.06
	(−0.4)	(5.8)	(−5.4)	(−0.8)	(−0.9)	(−1.0)	(0.4)	(0.4)
Illinois	—	—	—	—	—	—	—	—
Minnesota	0.25	−0.31	−0.29	−0.46	−0.34	−0.14	0.44	0.47
	(0.5)	(−2.2)	(−2.3)	(−2.1)	(−0.9)	(−0.4)	(2.0)	(2.2)
New York	−1.79	0.25	0.13	0.36	0.27	−0.09	−0.43	−0.35
	(−4.4)	(3.5)	(2.1)	(3.0)	(1.4)	(−0.5)	(−3.5)	(−2.9)
				AMI Patients				
California	−0.51	0.43	−0.22	−0.03	−0.18	−0.14	−0.04	−0.04
	(−1.2)	(6.1)	(−3.9)	(−0.3)	(−0.9)	(−0.7)	(−0.3)	(−0.3)
Illinois	—	—	—	—	—	—	—	—
Minnesota	−0.05	0.02	−0.12	−0.24	−0.21	−0.02	0.11	0.16
	(−0.1)	(0.2)	(−1.4)	(−1.7)	(−0.7)	(−0.1)	(0.7)	(1.0)
New York	−0.89	0.11	0.22	0.13	−0.04	−0.08	−0.15	−0.13
	(−2.1)	(1.6)	(4.2)	(1.3)	(−0.2)	(−0.4)	(−1.3)	(−1.1)

NOTE: t-statistics are in parentheses. Patients in Illinois constitute the omitted category. The independent variables used in Table 4 were also included in these equations; their estimated coefficients did not differ much from those in Table 4.

Table D.11

REGRESSION RESULTS FOR HOSPITAL CHARACTERISTICS

	Logit DNR	OLS Quality	OLS log(LOS) Alive	OLS log(LOS) Deadin	Logit Deadin	Logit Dead30	Cox Deadin	Cox Dead30
				CHF Patients				
Beds	0.00	0.00	−0.01	0.00	−0.05	−0.02	0.00	0.00
	(0.0)	(−0.1)	(−0.5)	(0.1)	(−1.2)	(−0.4)	(0.0)	(0.0)
Church	0.44	0.02	0.02	−0.13	−0.12	0.08	0.13	0.17
	(1.4)	(0.2)	(0.3)	(−1.1)	(−0.7)	(0.4)	(1.1)	(1.5)
Propri	−0.12	0.10	−0.22*	0.03	−0.17	−0.14	−0.05	−0.03
	(−0.2)	(1.0)	(−2.4)	(0.2)	(−0.6)	(−0.5)	(−0.3)	(−0.2)
Gvt	0.00	−0.07	−0.22*	−0.31*	−0.22	−0.09	0.34*	0.32*
	(0.0)	(−0.8)	(−2.7)	(−2.0)	(−0.9)	(−0.4)	(2.3)	(2.1)
Major	−0.13	0.08	0.14	0.18	−0.11	−0.14	−0.29	−0.30
	(−0.2)	(0.6)	(1.2)	(0.9)	(−0.3)	(−0.4)	(−1.4)	(−1.4)
Limited	0.42	−0.14	0.00	0.08	0.09	0.01	0.07	0.00
	(1.1)	(−1.5)	(0.0)	(0.5)	(0.4)	(0.0)	(0.5)	(0.0)
Graduate	−0.46	−0.13	−0.09	−0.01	−0.44	−0.22	0.06	0.00
	(−0.5)	(−0.8)	(−0.6)	(0.0)	(−1.0)	(−0.5)	(0.2)	(0.0)
Res_pgm	−0.48	0.03	0.10	0.12	0.32	0.09	0.01	−0.01
	(−1.0)	(0.3)	(0.9)	(0.7)	(1.2)	(0.3)	(0.0)	(0.0)
Rural	0.36	−0.50*	0.06	−0.11	0.03	0.26	0.10	0.16
	(0.9)	(−5.7)	(0.7)	(−0.8)	(0.1)	(1.1)	(0.7)	(1.2)
				AMI Patients				
Beds	−0.15	0.02	0.03*	−0.03	−0.02	−0.01	−0.02	−0.01
	(−1.3)	(1.5)	(2.3)	(−1.1)	(−0.5)	(−0.3)	(−0.8)	(−0.5)
Church	0.87*	−0.02	0.00	−0.11	−0.05	−0.03	−0.01	0.02
	(2.4)	(−0.3)	(−0.1)	(−1.2)	(−0.3)	(−0.2)	(0.0)	(0.2)
Propri	−0.01	0.06	−0.10	−0.20	−0.08	−0.03	−0.01	−0.02
	(0.0)	(0.6)	(−1.3)	(−1.3)	(−0.3)	(−0.1)	(0.0)	(−0.1)
Gvt	−2.10*	0.02	−0.19*	−0.10	−0.14	−0.05	0.06	0.08
	(−2.0)	(0.2)	(−2.9)	(−0.9)	(−0.6)	(−0.2)	(0.4)	(0.5)
Major	0.70	−0.14	0.02	−0.21	0.22	0.08	0.20	0.14
	(0.9)	(−1.2)	(0.3)	(−1.2)	(0.7)	(0.2)	(1.0)	(0.7)
Limited	0.40	−0.09	−0.01	0.05	0.08	0.01	0.07	0.05
	(0.7)	(−0.9)	(−0.1)	(0.4)	(0.3)	(0.1)	(0.5)	(0.3)
Graduate	0.38	−0.22	0.07	−0.22	−0.06	−0.16	0.12	0.05
	(0.4)	(−1.7)	(0.7)	(−1.1)	(−0.2)	(−0.4)	(0.5)	(0.2)
Res_pgm	−0.68	0.07	0.09	0.27	−0.20	−0.04	−0.25	−0.20
	(−1.0)	(0.8)	(1.2)	(1.8)	(−0.7)	(−0.2)	(−1.4)	(−1.1)
Rural	−0.11	−0.39*	−0.09	−0.16	−0.18	0.01	0.07	0.10
	(−0.2)	(−5.3)	(−1.4)	(−1.5)	(−0.9)	(0.0)	(0.6)	(0.8)

NOTES: t-statistics are in parentheses. Significant coefficients ($p < 0.05$) are marked with an asterisk to make them easier to pick out. Independent variables are: beds—number of certified beds; church—hospital is operated by a religious organization; propri—hospital is operated by a for profit organization; gvt—hospital is operated by a government organization; major—hospital has a major affiliation with a medical school; limited—hospital has a limited affiliation with a medical school; graduate—hospital has an affiliation with a graduate medical program; res_pgm—hospital has a residency program; and rural—hospital is located in a rural area. The independent variables used in Table 4 were also included in these equations; their estimated coefficients did not differ much from those in Table 4.

Appendix E

TARGETED HOSPITALS ARE SIMILAR TO UNTARGETED HOSPITALS

This appendix consists of counterparts to Tables 5 and 6 in the main text showing additional ways of comparing targeted and untargeted hospitals. The additional information included is described below.

Tables E.1 and E.2 add comparison of hospitals targeted by HCFA at the 0.05 level using 1986 data with those not so targeted (see Appendix A). (The comparisons show how these hospitals—targeted in 1986—differed in our 1984 data.) We use HCFA targeting for severe chronic heart disease to correspond to CHF, and HCFA targeting for severe acute heart disease to correspond to AMI. The correspondence is certainly not exact; the HCFA conditions are defined more broadly than ours (again see Appendix A). The comparison for HCFA targeting for chronic heart disease is based on only 26 sample patients in targeted hospitals.

Tables E.3 and E.4 use alternative ("ex ante") weights for the NOS 30-day and HCFA comparisons. In Tables E.1 and E.2, these comparisons use weights equal to the population to sample ratios in each of 16 cells that result from crossing the four inpatient sampling categories (untargeted-alive, untargeted-dead, targeted-alive, and targeted-dead) with the corresponding 30-day (or HCFA) categories. These 16-cell weights reflect the realized ("ex post") sampling proportions in each cell. But ex ante, our systematic random sample has *expected* sample proportions equal to the population proportions. Hence estimates using only the four-cell inpatient targeting population weights are ex ante unbiased and should not differ much from the 16-cell ex post weighted estimates. Tables E.3 and E.4 confirm this expectation.

Tables E.5 and E.6 disaggregate the summary quality of process scale into four of its subscales. Subscale differences between targeted and untargeted hospitals are spottily significant and go in the unexpected direction as often as not, in much the same way as the summary scale differences.

Tables E.7 and E.8 use less aggregated targeting classifications. Our original targeting method contrasts hospitals with $p < 0.05$ of having as many deaths as observed with all others. In Tables E.7 and E.8, we

compare four groups of hospitals, listed here from "best" to "worst":

Those with $p \geq 0.50$ of having as many deaths as observed;
Those with $p < 0.50$ but ≥ 0.05 of having as many deaths as observed;
Those with $p < 0.05$ but ≥ 0.01 of having as many deaths as observed;
Those with $p < 0.01$ of having as many deaths as observed.

Tables E.9 and E.10 are variants of Tables E.7 and E.8 that avoid a possible problem with the former tables. The problem is that some small hospitals with higher than expected deaths may be included in the "best" category because they have high p values simply because they treat so few patients. In Tables E.9 and E.10, the "best" category is defined as all hospitals with lower than expected death rates, regardless of p value. The "worst" hospitals are, as before, those with $p < 0.01$ of having as many deaths as observed. The middle category is all other hospitals.

Table E.11 tries to take advantage of purely random effects averaging out over time and so uses three years of data for targeting. Specifically, we multiply together the probabilities that a hospital would have as many deaths as it did from our 1984 30-day death analysis, and HCFA's 1988 analysis of 1986 and 1987 data. We then rank hospitals by the result of that computation and count as targeted that same number of hospitals from the top of the list that our 30-day method targeted in 1984.

Table E.12 looks for differences in quality of care received by patients who lived when expected to die ("miracles") and those who died when expected to live ("disasters"). We say that a patient is expected to die if the probability of death predicted based on severity score alone is greater than 0.50; a patient is expected to live if that predicted probability is less than 0.50.

Table E.1

SUMMARY STATISTICS FOR SAMPLED CHF PATIENTS IN FY 1984
BY SAMPLING CATEGORY

	Inpatient Targeting		30-Day Targeting		HCFA Targeting	
	Untargeted Hospitals	Targeted Hospitals	Untargeted Hospitals	Targeted Hospitals	Untargeted Hospitals	Targeted Hospitals
Patients in sample						
Alive	318	290	526	109	610	13
Dead	265	253	402	89	470	13
Total	583	543	928	198	1080	26
Severity score						
Alive	32.00	30.91	31.51	31.22	31.51	27.99 −
	(0.44)	(0.37)	(0.32)	(0.61)	(0.30)	(1.59)
Dead	41.37	39.75	−	40.39	40.37	40.28
	(0.56)	(0.56)	(0.45)	(0.88)	(0.40)	(2.82)
Weighted average	32.75	32.27	32.63	32.82	32.66	29.76
	(0.34)	(0.32)	(0.27)	(0.54)	(0.25)	(1.54)
DNR status at admission (%)						
Alive	2.20	0.34	− 1.02	0.00 −	0.97	0.00 −
	(0.82)	(0.34)	(0.44)	(0.00)	(0.40)	(0.00)
Dead	16.60	4.74 − −	13.34	23.22 +	16.06	9.29
	(2.29)	(1.34)	(1.70)	(4.50)	(1.70)	(8.38)
Weighted average	3.35	1.02 − −	2.57	4.08	2.94	1.21
	(0.75)	(0.43)	(0.52)	(1.41)	(0.51)	(2.18)
Quality of process score						
Alive	0.10	0.22	0.12	0.23	0.16	−0.49 ++
	(0.05)	(0.05)	(0.04)	(0.08)	(0.03)	(0.23)
Dead	−0.27	−0.14	−0.13	−0.42 +	−0.19	0.03
	(0.06)	(0.06)	(0.05)	(0.13)	(0.04)	(0.23)
Weighted average	0.07	0.16	0.09	0.11	0.11	−0.42 ++
	(0.04)	(0.04)	(0.03)	(0.07)	(0.03)	(0.16)
Length of stay (days)						
Alive	9.72	13.24 ++	10.60	13.28	10.55	27.30
	(0.39)	(1.29)	(0.39)	(3.04)	(0.35)	(19.81)
Dead	9.20	18.78 ++	8.84	8.41	8.86	6.42
	(0.58)	(1.42)	(0.32)	(0.71)	(0.30)	(1.54)
Weighted average	9.68	14.10 ++	10.38	12.42	10.33	24.59
	(0.29)	(0.95)	(0.28)	(2.06)	(0.26)	(12.88)

NOTES: Numbers tabulated are means with standard errors in parentheses. Inpatient results weighted by inverse inpatient sampling weights, others by respective 16-cell weights. Significant differences between untargeted and targeted hospitals are marked as follows: ++ $p < 0.01$, expected direction; + $p < 0.05$, expected direction; − $p < 0.05$, unexpected direction; and − − $p < 0.01$, unexpected direction.

Table E.2

SUMMARY STATISTICS FOR SAMPLED AMI PATIENTS IN FY 1984
BY SAMPLING CATEGORY

	Inpatient Targeting		30-Day Targeting		HCFA Targeting	
	Untargeted Hospitals	Targeted Hospitals	Untargeted Hospitals	Targeted Hospitals	Untargeted Hospitals	Targeted Hospitals
Patients in sample						
Alive	311	285	464	129	508	73
Dead	311	243	429	128	476	68
Total	622	528	893	257	984	141
Severity score						
Alive	21.63	20.92	21.61	22.68	21.65	20.28
	(0.65)	(0.65)	(0.53)	(0.96)	(0.50)	(1.53)
Dead	38.09	39.15	35.84	38.62	36.07	39.12
	(1.01)	(1.20)	(0.87)	(1.65)	(0.83)	(2.12)
Weighted average	24.92	26.44	24.91	28.11 ++	25.05	24.82
	(0.58)	(0.70)	(0.49)	(0.99)	(0.46)	(1.38)
DNR status at admission (%)						
Alive	0.96	0.00	0.95	0.00 –	1.00	0.00 –
	(0.56)	(0.00)	(0.45)	(0.00)	(0.44)	(0.00)
Dead	8.36	7.00	6.85	9.56	7.35	4.90
	(1.57)	(1.64)	(1.22)	(2.61)	(1.20)	(2.64)
Weighted average	2.45	2.12	2.32	3.26	2.50	1.21
	(0.62)	(0.63)	(0.50)	(1.11)	(0.50)	(0.92)
Quality of process score						
Alive	0.31	0.35	0.30	0.41	0.33	0.19
	(0.04)	(0.05)	(0.03)	(0.07)	(0.03)	(0.10)
Dead	0.05	0.05	0.12	0.15	0.12	0.14
	(0.06)	(0.06)	(0.05)	(0.08)	(0.05)	(0.12)
Weighted average	0.26	0.26	0.26	0.32	0.28	0.18
	(0.03)	(0.04)	(0.03)	(0.05)	(0.03)	(0.07)
Length of stay (days)						
Alive	13.17	15.78 ++	13.64	15.20	13.97	11.88 – –
	(0.35)	(0.65)	(0.33)	(1.09)	(0.34)	(0.66)
Dead	5.82	6.72	5.66	5.43	5.79	4.80
	(0.42)	(0.80)	(0.26)	(0.52)	(0.25)	(0.46)
Weighted average	11.70	13.05 +	11.80	11.87	12.04	10.14 – –
	(0.28)	(0.53)	(0.25)	(0.72)	(0.25)	(0.51)

NOTES: Numbers tabulated are means with standard errors in parentheses. Inpatient results weighted by inverse inpatient sampling weights, others by respective 16-cell weights. Significant differences between untargeted and targeted hospitals are marked as follows: ++ $p < 0.01$, expected direction; + $p < 0.05$, expected direction; – $p < 0.05$, unexpected direction; and – – $p < 0.01$, unexpected direction.

Table E.3

EFFECT OF EX ANTE INSTEAD OF EX POST WEIGHTING ON SUMMARY STATISTICS FOR SAMPLED CHF PATIENTS IN FY 1984 BY SAMPLING CATEGORY

	30-Day Targeting		HCFA Targeting	
	Untargeted Hospitals	Targeted Hospitals	Untargeted Hospitals	Targeted Hospitals
Patients in sample				
Alive	526	109	610	13
Dead	402	89	470	13
Total	928	198	1080	26
Severity score				
Alive	31.51	31.24	31.51	27.95 –
	(0.32)	(0.61)	(0.30)	(1.56)
Dead	40.43	40.14	40.28	41.30
	(0.45)	(0.83)	(0.40)	(2.81)
Weighted average	32.62	32.93	32.67	29.64 –
	(0.27)	(0.54)	(0.25)	(1.51)
DNR status at admission (%)				
Alive	1.01	0.00 –	0.96	0.00 –
	(0.44)	(0.00)	(0.39)	(0.00)
Dead	13.39	32.49 ++	16.16	10.46
	(1.70)	(4.99)	(1.70)	(8.83)
Weighted average	2.55	6.18 +	2.96	1.33
	(0.52)	(1.71)	(0.52)	(2.29)
Quality of process score				
Alive	0.12	0.23	0.16	−0.48 ++
	(0.04)	(0.08)	(0.03)	(0.23)
Dead	−0.13	−0.48 ++	−0.19	0.06
	(0.05)	(0.13)	(0.04)	(0.22)
Weighted average	0.09	0.09	0.11	−0.41 ++
	(0.03)	(0.07)	(0.03)	(0.16)
Length of stay (days)				
Alive	10.60	13.39	10.56	26.00
	(0.39)	(3.03)	(0.36)	(18.72)
Dead	8.83	8.47	8.85	6.27
	(0.33)	(0.68)	(0.30)	(1.48)
Weighted average	10.38	12.45	10.33	23.50
	(0.28)	(2.04)	(0.26)	(12.20)

NOTES: Numbers tabulated are means with standard errors in parentheses. All results weighted by inverse inpatient sampling weights. Significant differences between untargeted and targeted hospitals are marked as follows: ++ $p < 0.01$, expected direction; + $p < 0.05$, expected direction; – $p < 0.05$, unexpected direction; and – – $p < 0.01$, unexpected direction.

Table E.4

EFFECT OF EX ANTE INSTEAD OF EX POST WEIGHTING ON SUMMARY STATISTICS FOR SAMPLED AMI PATIENTS IN FY 1984 BY SAMPLING CATEGORY

	30-Day Targeting		HCFA Targeting	
	Untargeted Hospitals	Targeted Hospitals	Untargeted Hospitals	Targeted Hospitals
Patients in sample				
Alive	464	129	508	73
Dead	429	128	476	68
Total	893	257	984	141
Severity score				
Alive	21.61	22.02	21.65	20.28
	(0.53)	(1.02)	(0.50)	(1.53)
Dead	36.47	39.24	36.58	39.24
	(0.87)	(1.67)	(0.83)	(2.12)
Weighted average	24.88	28.40 ++	25.03	24.90
	(0.49)	(1.05)	(0.46)	(1.38)
DNR status at admission (%)				
Alive	0.95	0.00 –	1.00	0.00 –
	(0.45)	(0.00)	(0.44)	(0.00)
Dead	7.14	10.41	7.60	4.94
	(1.24)	(2.71)	(1.22)	(2.65)
Weighted average	2.32	3.86	2.50	1.23
	(0.50)	(1.20)	(0.50)	(0.93)
Quality of process score				
Alive	0.30	0.40	0.33	0.19
	(0.03)	(0.07)	(0.03)	(0.10)
Dead	0.10	0.12	0.10	0.14
	(0.05)	(0.08)	(0.05)	(0.12)
Weighted average	0.26	0.30	0.28	0.18
	(0.03)	(0.05)	(0.03)	(0.07)
Length of stay (days)				
Alive	13.62	15.98 +	13.96	11.88 – –
	(0.33)	(1.16)	(0.34)	(0.66)
Dead	5.53	5.17	5.65	4.70
	(0.26)	(0.52)	(0.25)	(0.45)
Weighted average	11.84	11.98	12.08	10.10 – –
	(0.25)	(0.76)	(0.25)	(0.51)

NOTES: Numbers tabulated are means with standard errors in parentheses. All results weighted by inverse inpatient sampling weights. Significant differences between untargeted and targeted hospitals are marked as follows: ++ $p < 0.01$, expected direction; + $p < 0.05$, expected direction; – $p < 0.05$, unexpected direction; and – – $p < 0.01$, unexpected direction.

Table E.5

SUMMARY STATISTICS FOR PROCESS SUBSCALES FOR SAMPLED CHF PATIENTS IN FY 1984 BY SAMPLING CATEGORY

	Inpatient Targeting		30-Day Targeting		HCFA Targeting	
	Untargeted Hospitals	Targeted Hospitals	Untargeted Hospitals	Targeted Hospitals	Untargeted Hospitals	Targeted Hospitals
Process – MD cognitive diagnostic						
Alive	0.10	0.38 – –	0.14	0.34 –	0.19	−0.72 ++
	(0.05)	(0.05)	(0.04)	(0.09)	(0.04)	(0.17)
Dead	−0.12	0.15 – –	0.00	−0.04	−0.03	0.27
	(0.06)	(0.06)	(0.05)	(0.10)	(0.04)	(0.24)
Weighted average	0.08	0.34 – –	0.12	0.27 –	0.16	−0.59 ++
	(0.04)	(0.04)	(0.03)	(0.06)	(0.03)	(0.14)
Process – RN cognitive diagnostic						
Alive	0.17	−0.01 +	0.14	0.10	0.14	0.22
	(0.05)	(0.06)	(0.04)	(0.07)	(0.04)	(0.16)
Dead	−0.13	−0.36 +	−0.08	−0.25	−0.10	−0.27
	(0.06)	(0.08)	(0.05)	(0.12)	(0.05)	(0.25)
Weighted average	0.15	−0.06 ++	0.12	0.04	0.11	0.16
	(0.04)	(0.04)	(0.03)	(0.06)	(0.03)	(0.13)
Process – technical diagnostic						
Alive	−0.09	0.12 – –	−0.06	0.09	−0.03	−0.79 +
	(0.05)	(0.05)	(0.04)	(0.10)	(0.04)	(0.32)
Dead	−0.33	−0.05 – –	−0.13	−0.42 +	−0.19	−0.24
	(0.08)	(0.06)	(0.05)	(0.14)	(0.05)	(0.24)
Weighted average	−0.11	0.10 – –	−0.07	0.00	−0.05	−0.72 ++
	(0.04)	(0.04)	(0.03)	(0.08)	(0.03)	(0.22)
Process – technical therapeutic						
Alive	0.06	−0.01	0.07	0.03	0.07	0.05
	(0.05)	(0.06)	(0.04)	(0.10)	(0.04)	(0.30)
Dead	−0.11	−0.07	−0.19	−0.24	−0.19	0.28 – –
	(0.07)	(0.07)	(0.06)	(0.15)	(0.06)	(0.16)
Weighted average	0.05	−0.02	0.04	−0.01	0.04	0.08
	(0.04)	(0.04)	(0.03)	(0.08)	(0.03)	(0.20)

NOTES: Numbers tabulated are means with standard errors in parentheses. Inpatient results weighted by inverse inpatient sampling weights, others by respective 16-cell weights. Significant differences between untargeted and targeted hospitals are marked as follows: ++ $p < 0.01$, expected direction; + $p < 0.05$, expected direction; – $p < 0.05$, unexpected direction; and – – $p < 0.01$, unexpected direction.

Table E.6

SUMMARY STATISTICS FOR PROCESS SUBSCALES FOR SAMPLED AMI PATIENTS IN FY 1984 BY SAMPLING CATEGORY

	Inpatient Targeting		30-Day Targeting		HCFA Targeting	
	Untargeted Hospitals	Targeted Hospitals	Untargeted Hospitals	Targeted Hospitals	Untargeted Hospitals	Targeted Hospitals
Process – MD cognitive diagnostic						
Alive	0.37	0.53 –	0.37	0.60 – –	0.38	0.34
	(0.05)	(0.05)	(0.04)	(0.07)	(0.04)	(0.10)
Dead	–0.05	0.00	0.04	0.06	0.03	0.25
	(0.06)	(0.07)	(0.05)	(0.09)	(0.05)	(0.12)
Weighted average	0.28	0.37	0.29	0.42 –	0.29	0.32
	(0.04)	(0.04)	(0.03)	(0.05)	(0.03)	(0.08)
Process – RN cognitive diagnostic						
Alive	0.14	–0.02 +	0.11	0.30 –	0.14	0.04
	(0.05)	(0.06)	(0.04)	(0.08)	(0.04)	(0.12)
Dead	–0.09	–0.13	–0.05	0.06	–0.04	–0.13
	(0.06)	(0.07)	(0.05)	(0.08)	(0.05)	(0.13)
Weighted average	0.10	–0.05 +	0.08	0.22 –	0.09	0.00
	(0.04)	(0.04)	(0.03)	(0.06)	(0.03)	(0.09)
Process – technical diagnostic						
Alive	0.16	0.34 – –	0.17	0.38 – –	0.19	0.04
	(0.04)	(0.04)	(0.04)	(0.06)	(0.03)	(0.09)
Dead	–0.11	–0.07	–0.04	–0.14	–0.04	0.06
	(0.07)	(0.07)	(0.06)	(0.10)	(0.05)	(0.12)
Weighted average	0.11	0.22 –	0.12	0.20	0.14	0.05
	(0.04)	(0.04)	(0.03)	(0.05)	(0.03)	(0.07)
Process – technical therapeutic						
Alive	0.07	0.10	0.05	–0.49 ++	0.06	–0.09
	(0.11)	(0.10)	(0.08)	(0.17)	(0.08)	(0.20)
Dead	0.23	0.07	0.25	0.28	0.28	–0.07
	(0.07)	(0.08)	(0.06)	(0.11)	(0.06)	(0.19)
Weighted average	0.12	0.09	0.12	–0.19 ++	0.13	–0.08
	(0.06)	(0.07)	(0.05)	(0.10)	(0.05)	(0.13)

NOTES: Numbers tabulated are means with standard errors in parentheses. Inpatient results weighted by inverse inpatient sampling weights, others by respective 16-cell weights. Significant differences between untargeted and targeted hospitals are marked as follows: ++ $p < 0.01$, expected direction; + $p < 0.05$, expected direction; – $p < 0.05$, unexpected direction; and – – $p < 0.01$, unexpected direction.

Table E.7

SUMMARY STATISTICS FOR SAMPLED CHF PATIENTS IN FY 1984
BY MORE DISAGGREGATED TARGETING CATEGORY

	Inpatient Probability				30-Day Probability			
	>0.50	<0.50	<0.05	<0.01	>0.50	<0.50	<0.05	<0.01
Patients in sample								
Alive	194	124	95	195	256	270	58	51
Dead	119	146	78	175	177	225	44	45
Total	313	270	173	370	433	495	102	96
Severity score								
Alive	32.41	31.36	30.53	31.09	31.38	31.72	31.58	30.76
	(0.58)	(0.66)	(0.65)	(0.45)	(0.47)	(0.44)	(0.73)	(1.03)
Dead	41.23	41.49	41.34	39.03 –	39.11	41.55	40.45	40.27
	(0.80)	(0.78)	(1.00)	(0.67)	(0.65)	(0.61)	(1.30)	(1.20)
Weighted average	32.94	32.46	32.11	32.34	32.15	33.33	33.10	32.48
	(0.47)	(0.50)	(0.59)	(0.38)	(0.38)	(0.38)	(0.69)	(0.85)
DNR status at admission (%)								
Alive	1.55	3.23	0.00	0.51	1.13	0.84	0.00	0.00
	(0.89)	(1.59)	(0.00)	(0.51)	(0.66)	(0.56)	(0.00)	(0.00)
Dead	16.81	16.44	10.26	2.29 – –	7.63	18.48	23.49	22.91 +
	(3.44)	(3.08)	(3.46)	(1.13)	(2.00)	(2.59)	(6.46)	(6.34)
Weighted average	2.46	4.66	1.50	0.79	1.78	3.73	4.02	4.15
	(0.88)	(1.29)	(0.93)	(0.46)	(0.64)	(0.85)	(1.95)	(2.05)
Quality of process score								
Alive	0.17	−0.01	0.14	0.26	0.12	0.12	0.36	0.06
	(0.06)	(0.08)	(0.09)	(0.06)	(0.06)	(0.05)	(0.11)	(0.13)
Dead	−0.25	−0.28	−0.29	−0.07	−0.01	−0.24	−0.29	−0.56 ++
	(0.09)	(0.09)	(0.10)	(0.08)	(0.06)	(0.06)	(0.16)	(0.19)
Weighted average	0.15	−0.04	0.08	0.21	0.11	0.07	0.25	−0.05
	(0.05)	(0.05)	(0.07)	(0.05)	(0.04)	(0.04)	(0.09)	(0.10)
Length of stay (days)								
Alive	9.17	10.58	12.02	13.84 +	10.28	11.04	15.90	10.03
	(0.46)	(0.68)	(1.06)	(1.85)	(0.50)	(0.60)	(5.51)	(1.12)
Dead	9.18	9.21	14.78	20.56 ++	8.35	9.29	9.67	6.95
	(0.77)	(0.85)	(2.10)	(1.82)	(0.49)	(0.43)	(1.09)	(0.87)
Weighted average	9.17	10.43	12.42	14.90 ++	10.10	10.76	14.83	9.47
	(0.37)	(0.48)	(0.91)	(1.33)	(0.38)	(0.42)	(3.79)	(0.79)

NOTES: Numbers tabulated are means with standard errors in parentheses. Inpatient results weighted by inverse inpatient sampling weights, others by respective 16-cell weights. Significant differences between hospitals with probability > 0.50 and probability < 0.01 of having as many deaths as they actually did are marked as follows: ++ $p < 0.01$, expected direction; + $p < 0.05$, expected direction; – $p < 0.05$, unexpected direction; and – – $p < 0.01$, unexpected direction.

Table E.8

SUMMARY STATISTICS FOR SAMPLED AMI PATIENTS IN FY 1984
BY MORE DISAGGREGATED TARGETING CATEGORY

	Inpatient Probability				30-Day Probability			
	>0.50	<0.50	<0.05	<0.01	>0.50	<0.50	<0.05	<0.01
Patients in sample								
Alive	196	115	164	121	223	241	82	47
Dead	149	162	138	105	196	233	80	48
Total	345	277	302	226	419	474	162	95
Severity score								
Alive	21.44	21.96	20.15	21.96	21.68	21.47	22.16	23.90
	(0.79)	(1.11)	(0.81)	(1.07)	(0.75)	(0.76)	(1.15)	(1.77)
Dead	38.22	37.96	40.25	37.71	34.60	37.82	35.87	44.07 ++
	(1.56)	(1.31)	(1.65)	(1.75)	(1.31)	(1.14)	(1.84)	(3.11)
Weighted average	24.11	26.13	26.18	26.79 +	24.44	25.84	26.67	31.27 ++
	(0.76)	(0.90)	(0.95)	(1.04)	(0.69)	(0.71)	(1.12)	(1.93)
DNR status at admission (%)								
Alive	1.53	0.00	0.00	0.00	1.37	0.06	0.00	0.00
	(0.88)	(0.00)	(0.00)	(0.00)	(0.78)	(0.16)	(0.00)	(0.00)
Dead	8.05	8.64	9.42	3.81	7.30	6.13	9.12	10.45
	(2.24)	(2.21)	(2.50)	(1.88)	(1.86)	(1.57)	(3.24)	(4.46)
Weighted average	2.57	2.25	2.82	1.17	2.64	1.68	3.00	3.82
	(0.85)	(0.89)	(0.95)	(0.72)	(0.78)	(0.59)	(1.35)	(1.98)
Quality of process score								
Alive	0.26	0.41	0.30	0.42	0.27	0.37	0.49	0.21
	(0.06)	(0.05)	(0.06)	(0.07)	(0.05)	(0.04)	(0.07)	(0.15)
Dead	0.06	0.04	−0.09	0.22	0.15	0.07	0.20	0.06
	(0.09)	(0.07)	(0.09)	(0.08)	(0.07)	(0.06)	(0.10)	(0.12)
Weighted average	0.23	0.32	0.18	0.36	0.25	0.29	0.40	0.15
	(0.05)	(0.04)	(0.05)	(0.05)	(0.04)	(0.03)	(0.06)	(0.10)
Length of stay (days)								
Alive	13.00	13.45	14.74	17.19 ++	13.57	13.79	14.23	17.45 ++
	(0.45)	(0.57)	(0.61)	(1.29)	(0.46)	(0.48)	(1.53)	(1.15)
Dead	5.75	5.89	7.47	5.74	5.46	5.99	5.91	4.49
	(0.60)	(0.59)	(1.25)	(0.85)	(0.35)	(0.39)	(0.71)	(0.66)
Weighted average	11.85	11.49	12.57	13.69	11.84	11.71	11.49	12.71
	(0.37)	(0.44)	(0.62)	(0.92)	(0.36)	(0.36)	(0.98)	(0.95)

NOTES: Numbers tabulated are means with standard errors in parentheses. Inpatient results weighted by inverse inpatient sampling weights, others by respective 16-cell weights. Significant differences between untargeted and targeted hospitals are marked as follows: ++ $p < 0.01$, expected direction; + $p < 0.05$, expected direction; − $p < 0.05$, unexpected direction; and − − $p < 0.01$, unexpected direction.

Table E.9

SUMMARY STATISTICS FOR SAMPLED CHF PATIENTS IN FY 1984 IN "BEST" COMPARED WITH "WORST" HOSPITALS

	Inpatient Deaths			30-Day Deaths		
	"Best"	"Middle"	"Worst"	"Best"	"Middle"	"Worst"
Patients in sample						
Alive	170	243	195	235	349	51
Dead	95	248	175	164	282	45
Total	265	491	370	399	631	96
Severity score						
Alive	32.30	31.50	31.09	31.25	31.83	30.76
	(0.62)	(0.47)	(0.45)	(0.48)	(0.39)	(1.03)
Dead	41.07	41.51	39.03	38.75	41.58	40.27
	(0.89)	(0.58)	(0.67)	(0.66)	(0.54)	(1.20)
Weighted average	32.78	32.63	32.34	32.02	33.34	32.48
	(0.51)	(0.37)	(0.38)	(0.39)	(0.33)	(0.85)
DNR status at admission (%)						
Alive	1.76	2.33	0.51	1.24	0.67	0.00
	(1.01)	(0.97)	(0.51)	(0.72)	(0.44)	(0.00)
Dead	16.84	15.29	2.29 ++	8.05	18.33	22.91 +
	(3.86)	(2.29)	(1.13)	(2.13)	(2.31)	(6.34)
Weighted average	2.59	3.78	0.79	1.94	3.40	4.15
	(0.98)	(0.86)	(0.46)	(0.69)	(0.72)	(2.05)
Quality of process score						
Alive	0.20	0.01	0.26	0.13	0.14	0.06
	(0.07)	(0.05)	(0.06)	(0.06)	(0.05)	(0.13)
Dead	−0.25	−0.28	−0.07	0.00	−0.24	−0.56 ++
	(0.10)	(0.07)	(0.08)	(0.07)	(0.06)	(0.19)
Weighted average	0.17	−0.02	0.21	0.12	0.08	−0.05
	(0.05)	(0.04)	(0.05)	(0.04)	(0.03)	(0.10)
Length of stay (days)						
Alive	8.82	10.93	13.84 ++	10.27	11.53	10.03
	(0.47)	(0.51)	(1.85)	(0.51)	(0.88)	(1.12)
Dead	9.20	10.25	20.56 ++	8.27	9.35	6.95
	(0.88)	(0.78)	(1.82)	(0.51)	(0.39)	(0.87)
Weighted average	8.84	10.85	14.90 ++	10.07	11.19	9.47
	(0.38)	(0.39)	(1.33)	(0.39)	(0.61)	(0.79)

NOTES: Numbers tabulated are means with standard errors in parentheses. Inpatient results weighted by inverse inpatient sampling weights, others by respective 16-cell weights. Significant differences between "best" and "worst" hospitals are marked as follows: ++ $p < 0.01$, expected direction; + $p < 0.05$, expected direction; − $p < 0.05$, unexpected direction; and − − $p < 0.01$, unexpected direction. "Best" hospitals are those with lower than expected death rates. "Worst" hospitals are those with $p < 0.01$ of having so many deaths (as in Tables E.7 and E.8). "Middle" hospitals are all others.

Table E.10

SUMMARY STATISTICS FOR SAMPLED AMI PATIENTS IN FY 1984 IN "BEST" COMPARED WITH "WORST" HOSPITALS

	Inpatient Deaths			30-Day Deaths		
	"Best"	"Middle"	"Worst"	"Best"	"Middle"	"Worst"
Patients in sample						
Alive	188	287	121	218	328	47
Dead	140	309	105	184	325	48
Total	328	596	226	402	653	95
Severity score						
Alive	21.35	21.81	21.96	21.45	21.96	23.90
	(0.81)	(0.69)	(1.07)	(0.76)	(0.65)	(1.77)
Dead	37.71	38.69	37.71	34.75	37.09	44.07 ++
	(1.61)	(0.98)	(1.75)	(1.35)	(0.96)	(3.11)
Weighted average	23.90	26.27	26.79 +	24.18	26.27	31.27 ++
	(0.77)	(0.63)	(1.04)	(0.70)	(0.60)	(1.93)
DNR status at admission (%)						
Alive	1.60	0.00	0.00	1.40	0.05	0.00
	(0.92)	(0.00)	(0.00)	(0.80)	(0.12)	(0.00)
Dead	7.14	9.37	3.81	7.39	6.51	10.45
	(2.18)	(1.66)	(1.88)	(1.93)	(1.37)	(4.46)
Weighted average	2.47	2.47	1.17	2.64	1.89	3.82
	(0.86)	(0.64)	(0.72)	(0.80)	(0.53)	(1.98)
Quality of process score						
Alive	0.28	0.36	0.42	0.27	0.38	0.21
	(0.06)	(0.04)	(0.07)	(0.05)	(0.04)	(0.15)
Dead	0.10	0.00	0.22	0.17	0.07	0.06
	(0.10)	(0.05)	(0.08)	(0.08)	(0.05)	(0.12)
Weighted average	0.25	0.26	0.36	0.25	0.29	0.15
	(0.05)	(0.03)	(0.05)	(0.04)	(0.03)	(0.10)
Length of stay (days)						
Alive	13.01	13.59	17.19 ++	13.58	13.81	17.45 ++
	(0.46)	(0.38)	(1.29)	(0.47)	(0.46)	(1.15)
Dead	5.58	6.26	5.74	5.59	5.78	4.49
	(0.53)	(0.54)	(0.85)	(0.37)	(0.33)	(0.66)
Weighted average	11.84	11.66	13.68	11.94	11.53	12.71
	(0.38)	(0.33)	(0.92)	(0.37)	(0.33)	(0.95)

NOTES: Numbers tabulated are means with standard errors in parentheses. Inpatient results weighted by inverse inpatient sampling weights, others by respective 16-cell weights. Significant differences between untargeted and targeted hospitals are marked as follows: ++ $p < 0.01$, expected direction; + $p < 0.05$, expected direction; − $p < 0.05$, unexpected direction; and − − $p < 0.01$, unexpected direction. "Best" hospitals are those with lower than expected death rates. "Worst" hospitals are those with $p < 0.01$ of having so many deaths (as in Tables E.7 and E.8). "Middle" hospitals are all others.

Table E.11

SUMMARY STATISTICS FOR SAMPLED CHF AND AMI PATIENTS IN FY 1984 BY SAMPLING CATEGORY WITH TARGETING BASED ON THREE YEARS OF DATA

	CHF Patients		AMI Patients	
	Untargeted Hospitals	Targeted Hospitals	Untargeted Hospitals	Targeted Hospitals
Patients in sample				
Alive	540	83	522	59
Dead	410	73	484	58
Total	950	156	1006	117
Severity score				
Alive	31.51	30.71	21.44	23.66
	(0.31)	(0.81)	(0.49)	(2.00)
Dead	40.28	40.43	36.78	38.03
	(0.43)	(1.00)	(0.82)	(2.27)
Weighted average	32.64	32.35	24.89	27.61
	(0.26)	(0.67)	(0.46)	(1.58)
DNR status at admission (%)				
Alive	0.69	3.86	0.97	0.00 –
	(0.36)	(2.13)	(0.43)	(0.00)
Dead	14.89	26.22 +	7.60	4.26
	(1.76)	(5.18)	(1.21)	(2.63)
Weighted average	2.51	7.63 +	2.46	1.21
	(0.51)	(2.13)	(0.49)	(1.00)
Quality of process score				
Alive	0.16	−0.10 ++	0.33	0.11
	(0.04)	(0.10)	(0.03)	(0.13)
Dead	−0.12	−0.67 ++	0.10	0.19
	(0.05)	(0.15)	(0.05)	(0.11)
Weighted average	0.13	−0.20 ++	0.28	0.13
	(0.03)	(0.08)	(0.03)	(0.08)
Length of stay (days)				
Alive	10.55	14.30	13.83	13.05
	(0.38)	(3.66)	(0.33)	(0.80)
Dead	8.99	7.23 –	5.63	4.80
	(0.32)	(0.71)	(0.25)	(0.57)
Weighted average	10.35	13.11	11.99	10.73
	(0.28)	(2.44)	(0.25)	(0.63)

NOTES: Numbers tabulated are means with standard errors in parentheses. All results weighted by respective 16-cell weights. Significant differences between untargeted and targeted hospitals are marked as follows: ++ $p < 0.01$, expected direction; + $p < 0.05$, expected direction; – $p < 0.05$, unexpected direction; and – – $p < 0.01$, unexpected direction.

Table E.12

SUMMARY STATISTICS FOR "MIRACLES"—CHF AND AMI SAMPLED PATIENTS WHO LIVED DESPITE A SEVERITY-PREDICTED PROBABILITY OF DYING > 0.5—AND "DISASTERS"—PATIENTS WHO DIED DESPITE A SEVERITY-PREDICTED PROBABILITY OF DYING < 0.5 IN FY 1984

	Inpatient Deaths		30-Day Deaths	
	Miracles	Disasters	Miracles	Disasters
CHF Patients				
Patients in sample				
Total	136	186	123	198
Severity score				
Weighted average	42.02	31.72 ++	42.67	32.14 ++
	(0.43)	(0.29)	(0.45)	(0.31)
DNR status at admission (%)				
Weighted average	5.45	5.85	1.54	5.91 −
	(1.95)	(1.72)	(1.11)	(1.68)
Quality of process score				
Weighted average	−0.08	−0.14	−0.08	0.10
	(0.08)	(0.08)	(0.08)	(0.07)
Length of stay (days)				
Weighted average	11.43	14.89 +	12.38	8.49 − −
	(0.61)	(1.36)	(0.88)	(0.47)
AMI Patients				
Patients in sample				
Total	125	199	128	203
Severity score				
Weighted average	38.63	20.38 ++	38.62	19.83 ++
	(0.69)	(0.46)	(0.68)	(0.45)
DNR status at admission (%)				
Weighted average	4.11	3.90	4.21	3.22
	(1.78)	(1.38)	(1.78)	(1.24)
Quality score				
Weighted average	0.23	0.40	0.23	0.44 −
	(0.08)	(0.06)	(0.08)	(0.06)
Length of stay (days)				
Weighted average	15.17	7.84 − −	15.65	7.20 − −
	(0.72)	(0.60)	(0.83)	(0.36)

NOTES: Numbers tabulated are means with standard errors in parentheses. Inpatient results weighted by inverse inpatient sampling weights, others by respective 16-cell weights. Significant differences between untargeted and targeted hospitals are marked as follows: ++ $p < 0.01$, expected direction; + $p < 0.05$, expected direction; − $p < 0.05$, unexpected direction; and − − $p < 0.01$, unexpected direction.

Appendix F

MORTALITY, LENGTH OF STAY, AND LOCATION OF DEATH: DISCUSSION AND GRAPHICAL COMPARISON OF INPATIENT AND 30-DAY DEATH MEASURES

We have previously discussed the relative advantages of inpatient compared with 30-day death measures as possible aids to identifying hospitals with potential quality problems (Chassin et al., 1989). We revisit these issues here with the help of some illuminating diagrams and some empirical results from this study. The curves in the diagrams plot the probability of dying on each day following admission to the hospital, for identical patients (i.e., with the same severity of illness), treated identically (i.e., receiving the same quality of care), both in the hospital and after discharge.[1]

HCFA shifted from inpatient to 30-day targeting after the first of its annual mortality analyses (the 1986 analysis of 1984 data; Brinkley, 1986). The discussion in this appendix generally supports that shift.

GRAPHICAL COMPARISON

Case 1

Figure F.1 shows the simplest (and least realistic) case. Patients have a declining probability of dying on each day following admission, independent of whether they are in hospital A, hospital B, or not in the hospital at all (top panel). The only difference is that hospital A discharges its patients at 13 days, and hospital B discharges at 20 days.

The inpatient death measure is the area under the probability curve out to the time of discharge (middle panel). Thus the difference

[1]The calculations described in Appendix G are based on related curves—so-called "hazard" curves—that show the risk of dying during each day for patients who were alive at the start of that day. The hazard curves differ from the probability curves used here in that the denominator for the hazard curve changes over time as patients die, and the denominator for the probability curves stays constant at the number of patients admitted.

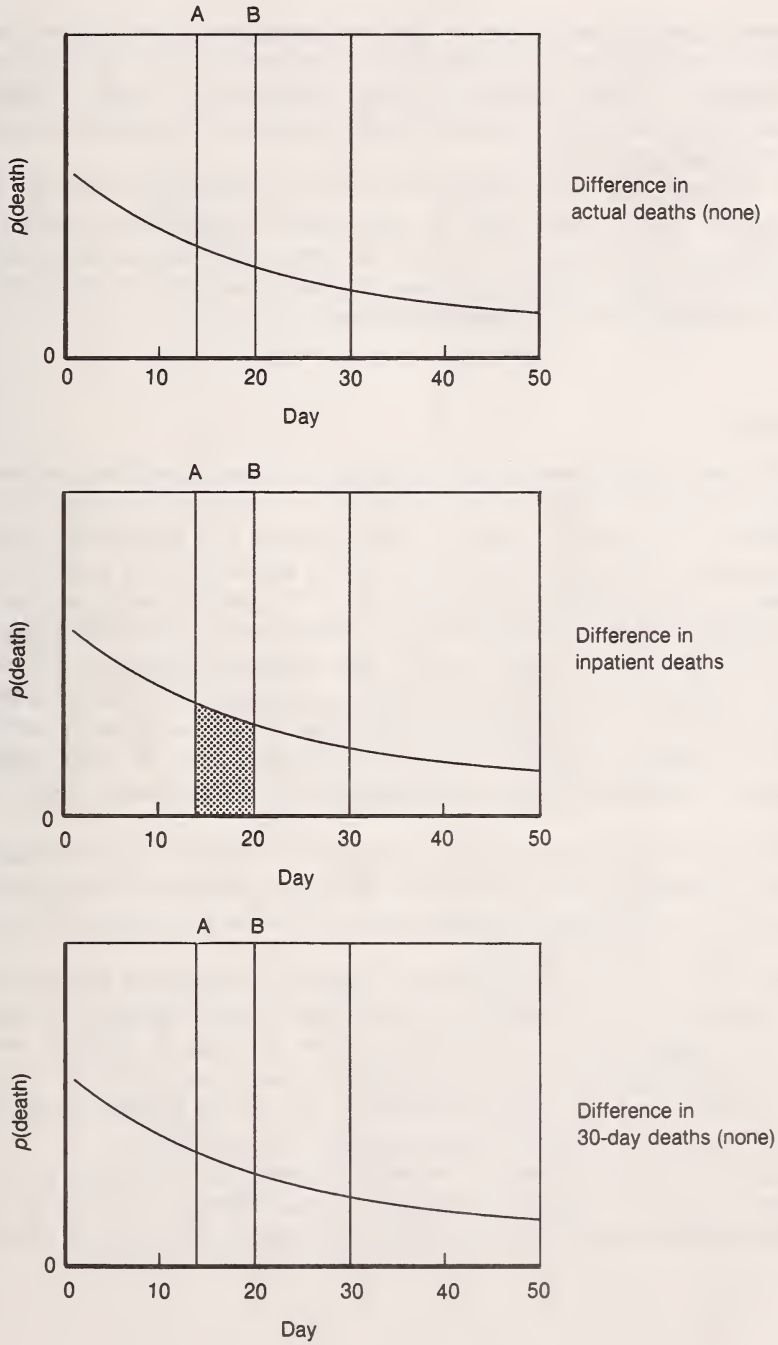

A
B
p(death)
0
0
10
20
30
40
50
Day
Difference in
actual deaths (none)
A
B
p(death)
0
0
10
20
30
40
50
Day
Difference in
inpatient deaths
A
B
p(death)
0
0
10
20
30
40
50
Day
Difference in
30-day deaths (none)

Fig. F.1—Case 1

between inpatient measures for the two hospitals is the shaded area under the probability curve between 13 days and 20 days. In this case, the inpatient death measure is clearly misleading, in that it differs between hospital A and hospital B, whereas there is no difference at all in actual patient outcomes.

In contrast, there is no difference in the 30-day death measures for the two hospitals, which equal the area under the probability curve out to 30 days (bottom panel). Thus in this case, the 30-day measure accurately reflects the lack of difference between actual patient outcomes in the two hospitals but the inpatient measure does not.

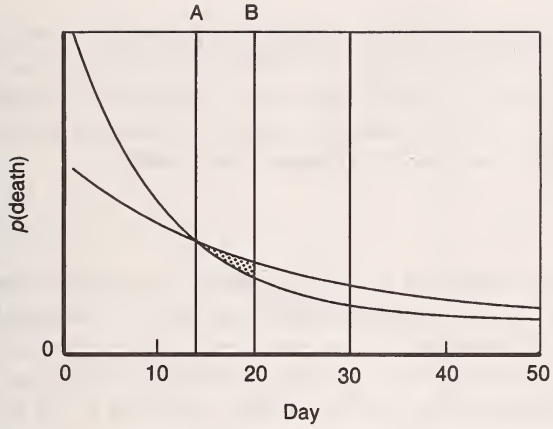
Case 2

We have argued previously that length of stay is itself a treatment decision, and too long (or too short) a stay can be bad care. This situation is illustrated in Fig. F.2, which shows two probability curves (top panel). The curve that is lower at the left side of the graph is for patients in the hospital; the higher one is for patients who are not in the hospital. The two curves cross at 13 days; up to that time, hospital care is beneficial, but after that, the risk of dying is higher in the hospital. Thus hospital A, which discharges its patients at 13 days, is providing optimal care, and hospital B, which keeps its patients for 20 days, is exposing them to additional risk after day 13. The actual amount of additional risk is measured by the shaded area in the top panel.

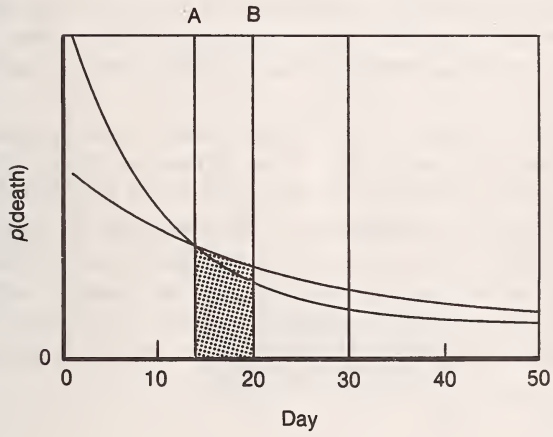
The inpatient death measure is larger for hospital B by the shaded area in the middle panel. This is the difference between the area under the curve for hospitalized patients out to 20 days (for hospital B) and out to 13 days (for hospital A).

The 30-day death measure is also larger for hospital B but only by the shaded area in the bottom panel. This is the difference between the area under the curve that shows how hospital B patients were treated and that showing how hospital A patients were treated, i.e., in the hospital until day 20 for hospital B, and in the hospital until day 13 for hospital A, with both areas extended out to 30 days.

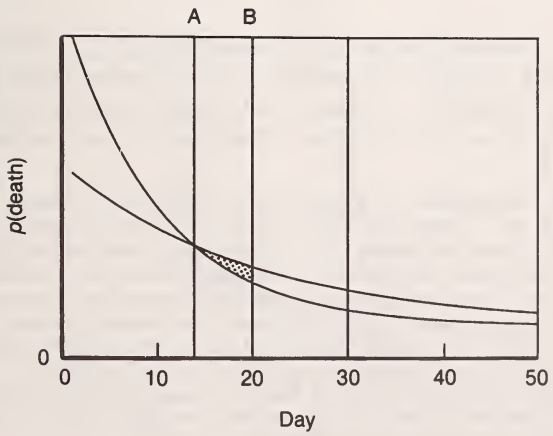
Again, the 30-day measure correctly reflects the difference in outcomes for the two hospitals. The only actual difference is during the period between day 13 and day 20, when hospital B's patients have a



Difference in actual deaths



Difference in inpatient deaths



Difference in 30-day deaths

Fig. F.2—Case 2

slightly higher risk of dying than do hospital A's patients.[2] The trouble with using inpatient deaths to measure differences between the hospitals is that it picks up not only the excess deaths in hospital B between day 13 and day 20, but also the deaths of patients receiving optimal care (i.e., outside the hospital) during that period.

Case 3

The situation changes some if hospital B keeps its patients longer than 30 days; Fig. F.3 shows a 40-day discharge date for hospital B. The actual difference in outcomes is the excess risk of dying because of being kept in hospital B from day 13 to day 40, or the shaded area in the top panel. The difference in inpatient death measures is the entire area under the inhospital probability curve from day 13 to day 40 (middle panel). The difference in 30-day death measures is the excess risk of dying because of being in the hospital from day 13 to day 30 (bottom panel).

In this situation neither inpatient nor 30-day deaths accurately measures the true difference in outcomes. The difference in inpatient deaths incorrectly includes patients who die outside the hospital from day 13 through day 40, and the difference in 30-day deaths incorrectly excludes the excess deaths from being in the hospital from day 30 to day 40. Which is the better measure is an empirical question with a situation-dependent answer, although unless the excess daily death rate is large relative to the absolute rate, the 30-day measure will be a better approximation.

Case 4

Of course, quality of care is not solely a matter of correct discharge decisionmaking. In our previous discussion, we emphasized that bad care can increase the probability of death and require longer hospital stays. Figure F.4 illustrates this situation. Hospital A treats patients

---

[2]We are finessing some complications in defining "actual differences" between outcomes with this simple statement. Actually, a higher probability of dying between days 13 and 20 must be offset by a lower probability of dying later, since everyone dies sometime. But it does matter *when* you die, with sooner usually judged to be worse than later. The differences in timing could be captured by using expected survival time (in the statistical sense of "expected") as an outcome measure but that is not commonly done. It is more common to use probability of survival for some fixed period (180 days, five years), or equivalently, death rates over some limited period. That is what we do here. The unstated assumption underlying the discussion of Fig. F.2, then, is that the out-of-hospital death probability curve is the same for all patients, regardless of prior treatment, at least until the end of whatever fixed period we have in mind.

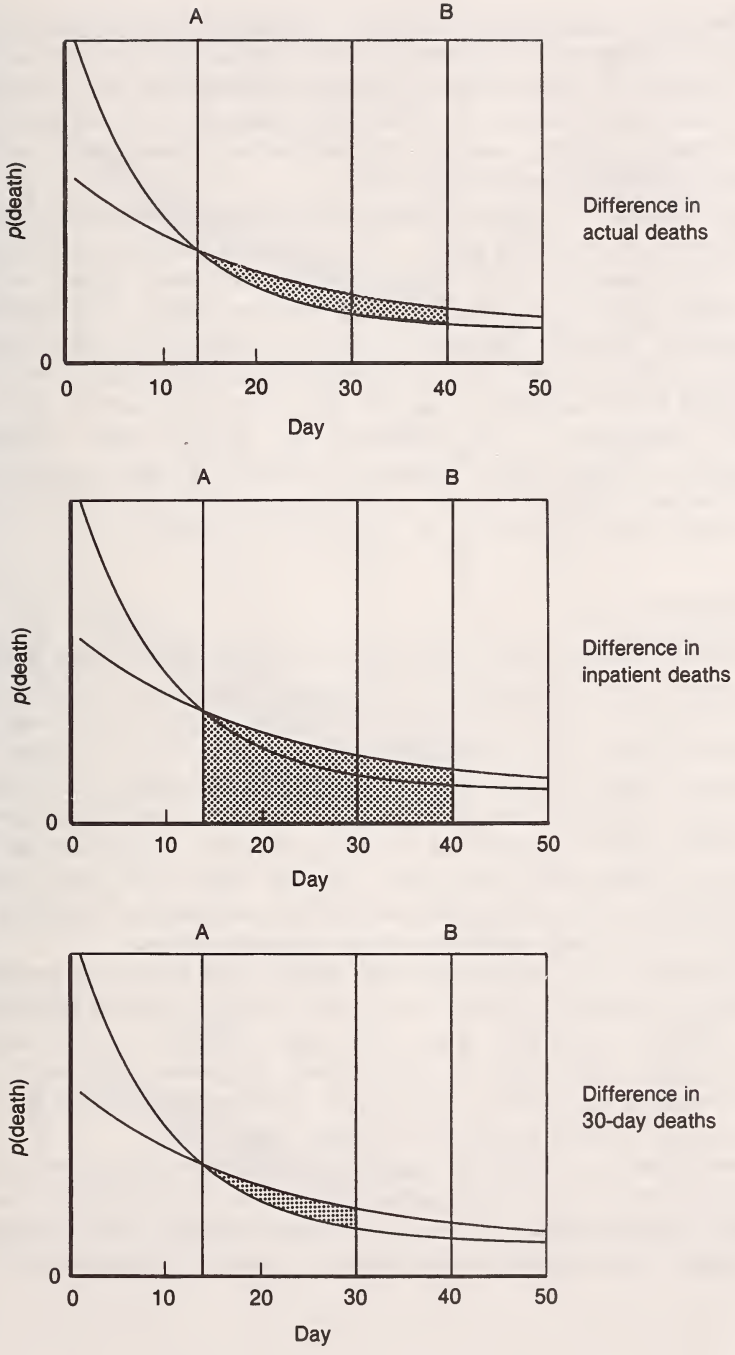

A
B
p(death)
0
0
10
20
30
40
50
Day
Difference in
actual deaths
A
B
p(death)
0
0
10
20
30
40
50
Day
Difference in
inpatient deaths
A
B
p(death)
0
0
10
20
30
40
50
Day
Difference in
30-day deaths

Fig. F.3—Case 3

as before, discharging them at 13 days. Its patients have a daily probability of death given by the bottom curve (top panel). Bad care in hospital B results in higher daily death probabilities (the top curve) and delays discharge until day 40. The true difference in outcomes is the area between the two curves. (For expository convenience, we are ignoring any differences after day 50.)

The difference in inpatient death measures is the saucepan shaped shaded area in the middle panel. The difference in 30-day death measures is the smaller shaded area in the bottom panel. Again, neither one exactly mirrors the actual difference in outcomes. The difference in inpatient measures incorrectly includes deaths from the time of discharge until 40 days after admission for patients who have been discharged from hospital A—deaths that would be netted out in a correct comparison. The difference in 30-day death measures incorrectly excludes the difference in hospital A's and hospital B's deaths between days 30 and 40. Both measures incorrectly exclude any differences in deaths after day 40 (up to day 50 in the figure).

Assessment

The main problem with using inpatient deaths to compare hospital outcomes is that the comparison charges the longer-stay hospital with *all* of its deaths between the shorter and longer lengths of stay, not just its excess deaths. On this account the comparison overstates the difference between longer-stay and shorter-stay hospitals. That problem can be solved by comparing deaths over any fixed-length period. The problem with using 30-day deaths is that the comparison fails to include real differences that occur after the end of the fixed period. That problem can in principle be solved by extending the length of the fixed period to include all effects of the hospitalizations.

In practice, very long fixed-period deaths have their own problems as a way to compare hospital outcomes, because patients discharged from different hospitals may face very different out-of-hospital environments. The longer the fixed period, the greater the chance that the different environments will result in different death rates for reasons that are entirely unrelated to hospital care.

Nevertheless, our diagrams suggest that a search for the best single death measure to use in identifying hospitals with potential quality problems should better be focused on deaths during shorter compared with longer fixed periods, rather than on inpatient compared with 30-

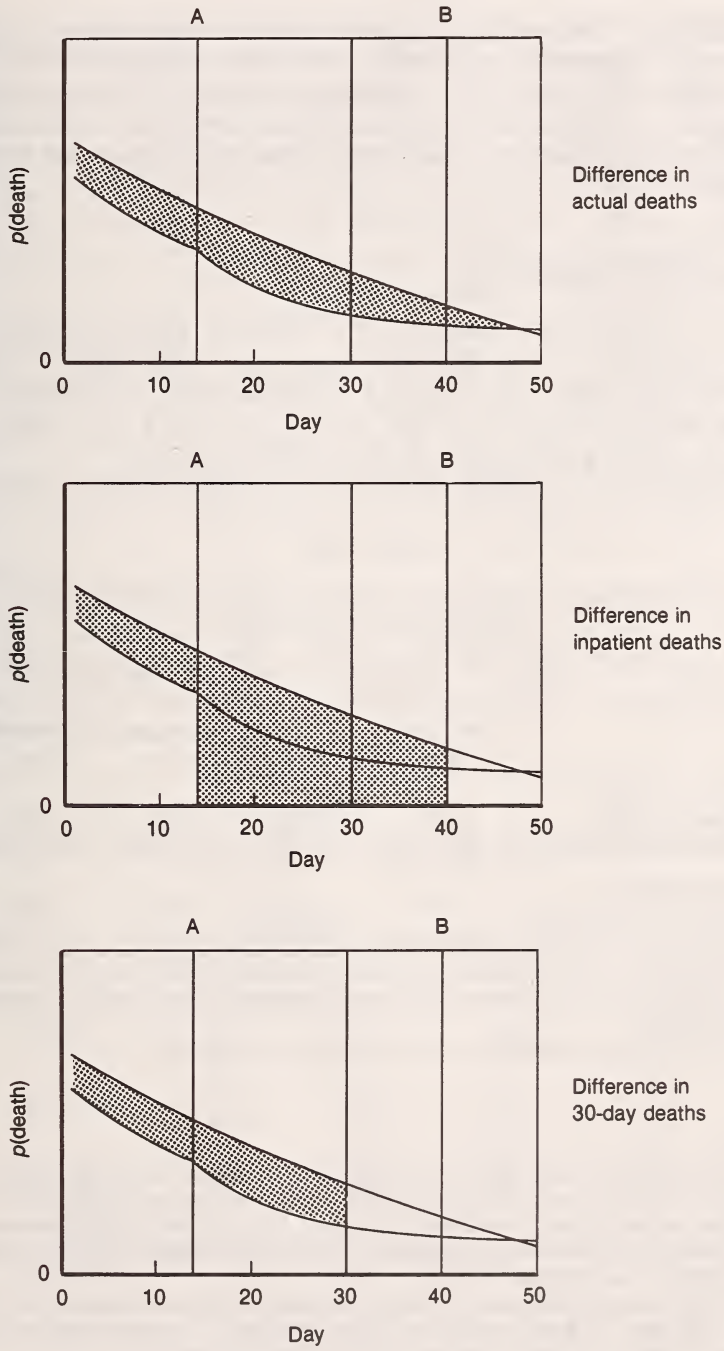

A
B
p(death)
0
0
10
20
30
40
50
Day
Difference in
actual deaths
A
B
p(death)
0
0
10
20
30
40
50
Day
Difference in
inpatient deaths
A
B
p(death)
0
0
10
20
30
40
50
Day
Difference in
30-day deaths

Fig. F.4—Case 4

day deaths.[3] If the search is not restricted to a single measure, then other ways of summarizing hospital death rates can provide addition useful information. Any single measure can conceal differences in the timing (day 1 after admission or day 10) and location of death (in the hospital or after discharge) which may correspond to important clinical differences.

EMPIRICAL EVIDENCE

All of the empirical evidence in this section is given separately for patients discharged alive and discharged dead. The reason is that length of stay means quite different things for those two groups. For patients who live, length of stay is the result of a treatment and discharge planning decision; for patients who die, it is not.[4]

Table F.1 shows differences in average lengths of stay between

Table F.1

DIFFERENCES IN LENGTH OF STAY BETWEEN TARGETED AND UNTARGETED HOSPITALS

	CHF Patients		AMI Patients	
	Discharged Alive	Discharged Dead	Discharged Alive	Discharged Dead
Inpatient targeting				
Difference (days)	3.52	9.58	2.62	0.90
95% confidence interval	(0.89, 6.16)	(6.57,12.60)	(1.16, 4.07)	(−0.87, 2.67)
30-day targeting				
Difference (days)	2.29	0.01	2.37	−0.80
95% confidence interval	(−3.58, 8.15)	(−3.44, 3.47)	(0.17, 4.57)	(−2.45, 0.84)

SOURCES: Tables 5 and 6 for inpatient targeting; 30-day targeting differs from Tables 5 and 6 because here it is cross-classified by inpatient death, and there by 30-day death.

NOTE: Positive difference is longer stay in targeted hospitals.

[3]In Chassin et al. (1989), one measure we used was deaths during the period from admission to the 95th percentile length of stay, which is one way to accommodate varying average lengths of stay for different conditions. Many variants of that measure are possible and may be attractive, for example, 95th percentile length of stay for patients discharged alive plus 10 days.

[4]To relate this distinction back to the previous section, note that length of stay in the diagrams (13 days, 20 days, 40 days) is of the former kind; it is when patients will be discharged if they live. Patients may die (and be discharged) earlier.

targeted and untargeted hospitals (as targeted minus untargeted, so that positive numbers indicate longer stays in targeted hospitals), together with 95 percent confidence bounds on the differences. Patients discharged alive stay on average two or three days longer in targeted hospitals; this is true for both inpatient targeting and 30-day targeting, and for both CHF and AMI (although the difference is not statistically significant for 30-day targeting for CHF). There is little or no difference in average length of stay for patients discharged dead, with the striking exception of CHF patients who survived more than nine days longer on average in inpatient targeted hospitals than in untargeted hospitals.

By themselves, these comparisons do not tell us anything about the reasons for longer stays in targeted hospitals. The differences could in principle be caused by differences in case mix, quality of care, discharge constraints such as lack of nursing home beds, or customary practice styles. We conjectured in our earlier discussion that sicker patients would tend to have longer stays and that bad care could cause longer stays.

The regression results in Table F.2 shed light on some of these relationships. Severity of illness does have the conjectured effect on length of stay for patients who get out of the hospital alive; sicker ones stay longer. For those who die in the hospital, it is not too surprising to find that sicker ones die faster.

Table F.2

ORDINARY LEAST SQUARES REGRESSION RESULTS FOR THE LOGARITHM OF LENGTH OF STAY

	CHF Patients		AMI Patients	
	Discharged Alive	Discharged Dead	Discharged Alive	Discharged Dead
Severity	1.11	−2.14	0.48	−2.10
	(3.1)	(−4.2)	(2.6)	(−9.9)
Quality	−0.01	0.14	0.05	0.17
	(−0.5)	(3.2)	(2.0)	(4.1)
R-square	0.02	0.11	0.02	0.22
Observations	608	518	596	553

SOURCE: Table 4.

NOTE: Estimates for the constant term and the effect of DNR are omitted from this table; t-statistics are in parentheses.

Our conjectured relation of bad care with long stays is not supported by these data. For patients discharged alive, there is little relationship one way or the other, with some suggestion that for AMI patients, *better* care is associated with longer stays. For patients who die in the hospital, it is clearly true that better care keeps them alive longer. Similar relationships are apparent in the bivariate correlations in Table F.3.

In summary, the empirical results do not give us any reason to reverse the conclusion that fixed-period death measures are likely to be superior to inpatient deaths in most situations.

Table F.3

CORRELATIONS OF THE LOGARITHM OF LENGTH OF STAY WITH SEVERITY AND QUALITY

	Discharged Alive			Discharged Dead		
	log(LOS)	Severity	Quality	log(LOS)	Severity	Quality
CHF patients		(n = 608)			(n = 518)	
log(LOS)	1.0000			1.0000		
Severity	0.1270	1.0000		−0.2215	1.0000	
Quality	−0.0251	−0.0418	1.0000	0.1878	−0.1116	1.0000
AMI patients		(n = 596)			(n = 553)	
log(LOS)	1.0000			1.0000		
Severity	0.0999	1.0000		−0.4413	1.0000	
Quality	0.0790	−0.0524	1.0000	0.2850	−0.2836	1.0000

Appendix G

"EXPLAINING" OBSERVED DIFFERENCES IN DEATH RATES BETWEEN TARGETED AND UNTARGETED HOSPITALS

This appendix describes how we calculated the effect on death rates of systematic differences between targeted and untargeted hospitals—effects that are summarized in Tables 7 and 8 in the main text.

We started by calculating 95 percent confidence limits for the estimated differences in severity, DNR, quality, and length of stay. These calculations are straightforward, given the data that underlie text Tables 5 and 6. The resulting confidence intervals are collected in Tables G.1 and G.2. (Throughout this appendix, we show numbers to three decimal places for two reasons: to facilitate unambiguous reference in the text to numbers in the tables and to support accurate rounding to the numbers used in the text tables.)

To translate the differences in the independent variables (in Tables G.1 and G.2) into differences in death rates (in Table G.4), we relied on the Cox model estimates reported in columns (7) and (8) of Table 4 in the main text. The coefficients, b, in Table 4, together with the values of the independent variables for a particular individual, X_i, determine the *relative* hazard of that individual dying on any particular day after admission, $\exp(X_i\ b)$, a value that differs from patient to patient, but for a given patient is constant over time. The Cox model also estimates a *baseline* hazard, H_t, that varies over time but is the same for all patients. An individual patient's risk of dying on a particular day is predicted as $H_{it} = H_t\ \exp(X_i\ b)$.

Corresponding to the baseline hazard is a baseline survival function, S_t, which can be calculated by repeated application of $S_t = S_{t-1}\ (1 - H_t)$. An individual patient's predicted probability of surviving through day t is $S_t \hat{}\exp(X_i\ b)$, where $\hat{}$ indicates exponentiation.

We used the estimated Cox models to make a number of "what if" predictions of death rates as described below. First, though, we had to adjust the estimated baseline hazard and survival functions, because the estimates are unweighted (we could not find software that supported weighted estimation). Therefore, the baseline hazard reflects sample death rates, which are higher than population death rates

Table G.1

ESTIMATED DIFFERENCES IN SEVERITY, DNR, QUALITY, AND LENGTH OF STAY FOR CHF PATIENTS

	Untargeted Hospitals	Targeted Hospitals	Difference
Targeting Using Inpatient Deaths			
Severity score	32.747	32.268	−0.480
95% confidence interval	(32.073, 33.421)	(31.640, 32.895)	(−1.400, 0.441)
DNR	3.346	1.022	−2.324
95% confidence interval	(1.885, 4.807)	(0.175, 1.869)	(−4.012, −0.635)
Quality score	0.073	0.164	0.091
95% confidence interval	(0.002, 0.143)	(0.088, 0.240)	(−0.013, 0.195)
Length of stay	9.678	14.097	4.418
95% confidence interval	(9.102, 10.255)	(12.233, 15.960)	(2.468, 6.369)
Targeting Using 30-Day Deaths			
Severity score	32.631	32.825	0.194
95% confidence interval	(32.103, 33.158)	(31.767, 33.883)	(−0.988, 1.376)
DNR	2.570	4.079	1.508
95% confidence interval	(1.551, 3.589)	(1.317, 6.841)	(−1.436, 4.452)
Quality score	0.091	0.114	0.023
95% confidence interval	(0.035, 0.147)	(−0.019, 0.247)	(−0.122, 0.167)
Length of stay	10.380	12.424	2.044
95% confidence interval	(9.827, 10.933)	(8.393, 16.455)	(−2.024, 6.113)

because we oversampled deaths. Thus we adjusted the baseline hazard downward by trial and error (in the same proportion every day), calculating at each trial the implied higher survival function, until the weighted sum of predicted death probabilities equaled the actual overall population death rate. We used the Cox models with the resulting adjusted baseline hazard and survival functions to make the "what if" predictions.

Table G.3 shows some of the "what if" predictions for AMI 30-day deaths to illustrate the method. The population weighted average probability of death, predicted using actual values for all of the independent variables, is 23.657 percent in untargeted hospitals and 26.975 percent in targeted hospitals. (The predicted probabilities for individual patients were calculated as 1 minus the predicted probability of surviving through day 30, $1 - S_{30}\hat{}\exp(X_i\ b)$.) Thus the predicted spread, at actual values of the independent variables, between death

Table G.2

ESTIMATED DIFFERENCES IN SEVERITY, DNR, QUALITY, AND LENGTH OF STAY FOR AMI PATIENTS

	Untargeted Hospitals	Targeted Hospitals	Difference
Targeting Using Inpatient Deaths			
Severity score	24.918	26.438	1.520
95% confidence interval	(23.778, 26.058)	(25.061, 27.814)	(−0.267, 3.307)
DNR	2.445	2.117	−0.328
95% confidence interval	(1.230, 3.660)	(0.888, 3.346)	(−2.056, 1.400)
Quality score	0.262	0.256	−0.006
95% confidence interval	(0.198, 0.326)	(0.184, 0.328)	(−0.102, 0.090)
Length of stay	11.697	13.041	1.344
95% confidence interval	(11.139, 12.255)	(12.001, 14.081)	(0.164, 2.524)
Targeting Using 30-Day Deaths			
Severity score	24.906	28.112	3.206
95% confidence interval	(23.953, 25.859)	(26.165, 30.059)	(1.038, 5.374)
DNR	2.319	3.258	0.938
95% confidence interval	(1.332, 3.307)	(1.083, 5.432)	(−1.450, 3.327)
Quality score	0.262	0.320	0.059
95% confidence interval	(0.208, 0.315)	(0.220, 0.421)	(−0.055, 0.173)
Length of stay	11.791	11.874	0.083
95% confidence interval	(11.299, 12.284)	(10.459, 13.290)	(−1.416, 1.582)

Table G.3

ILLUSTRATIVE COMPARISONS OF PREDICTED DEATH RATES USING ACTUAL AND HYPOTHETICAL LEVELS OF SEVERITY FOR AMI PATIENTS
(Deaths per 100 admissions)

	30-Day Untargeted Hospitals	30-Day Targeted Hospitals	Spread
Using actual values for all independent variables	23.657	26.975	3.318
Subtracting 3.2206 from severity scores for all patients in targeted hospitals, so that there is no difference in average severity scores	23.657	24.163	0.506
Subtracting 2.168 = 3.206 − 1.038 from severity scores for all patients in targeted hospitals, so that the difference in average severity scores is the lowest plausible value	23.657	25.048	1.391
Adding 2.168 = 5.374 − 3.206 to severity scores for all patients in targeted hospitals, so that the difference in average severity scores is the highest plausible value	23.657	29.009	5.352

rates in targeted and untargeted hospitals, is 26.975 – 23.657 = 3.318 deaths per 100 admissions.

The first "what if" or counterfactual prediction in Table G.3 reduces severity scores for all patients in targeted hospitals by enough so that the average score is the same in both targeted and untargeted hospitals. From Tables G.1 and G.2, one way to equate the average severity scores is to reduce them by 3.206 for every patient in targeted hospitals. We did that and recalculated predicted death rates. The population weighted average predicted probability of death dropped to 24.163 percent in targeted hospitals, reducing the spread between the predictions for targeted and untargeted hospitals to 0.506. We take that to mean that severity differences account for 3.318 – 0.506 = 2.812 of the difference in deaths per 100 admissions between targeted and untargeted hospitals. This number, 2.812, appears in Table G.4 as the predicted effect of actual severity differences on the difference in AMI 30-day death rates between targeted and untargeted hospitals.

So far we have described how we calculated the effect of severity differences at the center of the estimated confidence region for severity differences. In addition, we calculated the effect of severity differences at the lower and upper bounds of that confidence region. To do so, we adjusted severity scores for all patients in targeted hospitals so that the population weighted average severity score in targeted hospitals exceeded that in untargeted hospitals by the lower or upper bound amount and then did the "what if" calculations. For example, Tables G.1 and G.2 show that the lower confidence bound for severity differences between AMI 30-day targeted and untargeted hospitals was 1.1038. One way to get that size of a difference is to subtract 2.168 from the severity scores for all patients in targeted hospitals. (2.168 is the distance between the center of the confidence interval, 3.206, and the lower bound, 1.038.) At this lower level of severity, the predicted average death rate in targeted hospitals fell to 25.048, leaving a predicted spread of 25.048 – 23.657 = 1.391. This compares with a spread of 0.506 previously calculated for no severity difference. We take that to mean that an average severity difference at the lower bound of the confidence interval accounts for 1.391 – 0.506 = 0.885 of the difference in deaths per 100 admissions between targeted and untargeted hospitals. This number, 0.885, appears in Table G.4 as the predicted effect of lower bound severity differences. The effects of upper bound differences were similarly calculated.

The "what if" calculations for inpatient deaths differ in two ways from the calculations just described for 30-day deaths. Both differences arise because the observations for the Cox inpatient estimates were truncated at discharge from the hospital rather than uniformly at

Table G.4

DIFFERENCES IN DEATH RATE CORRESPONDING TO ESTIMATED DIFFERENCES IN SEVERITY, QUALITY, DNR, AND LENGTH OF STAY
(Deaths per 100 admissions)

Explanatory Variable	Inpatient Deaths[a]	30-Day Deaths[a]
	CHF Patients	
Due to severity difference	−0.252	0.160
95% confidence interval	(−0.723, 0.237)	(−0.790, 1.175)
Due to DNR difference	−0.194	0.165
95% confidence interval	(−0.332, −0.053)	(−0.155, 0.491)
Due to quality difference	−0.107	−0.036
95% confidence interval	(0.015, −0.229)	(0.190, −0.258)
Due to length of stay difference	3.174	—
95% confidence interval	(1.789, 4.520)	
	AMI Patients	
Due to severity difference	1.202	2.812
95% confidence interval	(−0.206, 2.687)	(0.885, 4.846)
Due to DNR difference	−0.020	0.080
95% confidence interval	(−0.125, 0.085)	(−0.123, 0.285)
Due to quality difference	0.017	−0.124
95% confidence interval	(0.282, −0.246)	(0.116, −0.362)
Due to length of stay difference	1.133	—
95% confidence interval	(0.139, 2.104)	

[a]A negative sign means that if actual differences in severity, DNR status at admission, or quality were eliminated between targeted (high mortality) hospitals and untargeted hospitals, death rates at targeted hospitals would increase.

30 days following admission. The first difference is that the predicted probabilities for individual inhospital deaths were calculated as 1 minus the predicted probability of surviving through discharge, $1 - S_t\hat{}\exp(X_i\ b)$, where t is a discharge day, which varies from patient to patient. The second difference is that we did an additional "what if" calculation for inpatient deaths. Besides predicting deaths if severity, DNR, or quality differed from observed values, we also predicted deaths if length of stay differed from observed values.

For example, we estimated average length of stay for CHF patients in inpatient targeted hospitals to be 14.097 days and in untargeted hospitals to be 9.678 days (Tables G.1 and G.2). To do the "what if" calculation for no difference in length of stay between targeted and untargeted hospitals, we multiplied each patient's length of stay in a

targeted hospital by the fraction 9.678/14.091. We then predicted probability of death at the earlier discharge dates (where the estimated survival curve was higher), and averaged the resulting predictions (Table G.5). That reduced the predicted spread in death rates from 1.860 (using actual length of stay) to −1.314 (using hypothetical length of stay). We take that to mean that length of stay differences account for 1.860 − (−1.314) = 3.174 of the difference in deaths per 100 admissions between targeted and untargeted hospitals. This number, 3.174, appears in Table G.4 as the predicted effect of actual length of stay differences on the difference in CHF inpatient death rates between targeted and untargeted hospitals.

Table G.5

ILLUSTRATIVE COMPARISONS OF PREDICTED DEATH RATES USING ACTUAL AND HYPOTHETICAL LEVELS OF LENGTH OF STAY FOR CHF PATIENTS
(Deaths per 100 admissions)

	Inpatient Untargeted Hospitals	Inpatient Targeted Hospitals	Spread
Using actual values for all independent variables	9.029	10.889	1.860
Multiplying length of stay by 9.678/14.091 for all patients in targeted hospitals, so that there is no difference in average length of stay	9.029	7.715	−1.314

BIBLIOGRAPHY

Berwick, D. M., and D. L. Wald, "Hospital Leaders' Opinions of the HCFA Mortality Data," *JAMA,* Vol. 263, No. 2, 1990, pp. 247–249.

Blumberg, M. S., "Comments on HCFA Hospital Death Rate Statistical Outliers," *Health Services Research,* Vol. 21, No. 6, 1987, pp. 715–740.

Bowen, O. R., and W. L. Roper, *Medicare Hospital Mortality Information, 1986,* Health Care Financing Administration, publication number 01-002, U.S. Department of Health and Human Services, 1987.

Bowen, O. R., and W. L. Roper, *Medicare Hospital Mortality Information, 1987,* Health Care Financing Administration, publication number 00651, U.S. Department of Health and Human Services, 1988.

Brinkley, J., "U.S. Releasing Lists of Hospitals with Abnormal Mortality Rates," *The New York Times,* March 12, 1986, p. 1.

Bunker, J. P., W. H. Forrest, Jr., F. Mosteller, et al., *The National Halothane Study,* U.S. Government Printing Office, Washington, D.C., 1969.

Chassin, M. R., R. E. Park, K. N. Lohr, et al., "Differences among Hospitals in Medicare Patient Mortality," *Health Services Research,* Vol. 24, No. 1, 1989, pp. 1–31.

Codman, E. A., "The Product of a Hospital," *Surg. Gynecol. Obstet.,* April 1914, pp. 491–496.

Department of Health and Human Services, "Final Regulations on Acquisition, Protection, and Disclosure of Peer Review Information," *Federal Register,* Vol. 50, No. 74, April 17, 1985, pp. 11347–11364.

Dubois, R. W., R. H. Brook, W. H. Rogers, "Adjusted Hospital Death Rates: A Potential Screen for Quality of Medical Care," *Am. J. Public Health,* Vol. 77, 1987, pp. 1162–1166.

Dubois, R. W., W. H. Rogers, J. H. Moxley, et al., "Hospital Inpatient Mortality: Is It a Predictor of Quality?" *New Engl. J. Med.,* Vol. 317, 1987, pp. 1674–1680.

Farber, B. F., D. L. Kaiser, and R. P. Wenzel, "Relation between Surgical Volume and Incidence of Postoperative Wound Infection," *New Engl. J. Med.,* Vol. 305, 1981, pp. 200–204.

Fink, A., E. M. Yano, and R. H. Brook, "The Condition of the Literature on Differences in Hospital Mortality," *Med. Care,* Vol. 27, No. 4, 1989, pp. 315–336.

Flood, A. B., and W. R. Scott, *Hospital Structure and Performance,* Johns Hopkins Press, Baltimore, 1987.

Flood, A. B., W. Ewy, W. R. Scott, et al., "The Relationship between Intensity and Duration of Medical Services and Outcomes for Hospitalized Patients," *Med. Care,* Vol. 17, 1979, pp. 1088–1102.

Flood, A. B., W. R. Scott, and W. Ewy, "Does Practice Make Perfect? Part I: The Relation between Hospital Volume and Outcomes for Selected Diagnostic Categories," and "Part II: The Relation between Volume and Outcomes and Other Hospital Characteristics," *Med. Care,* Vol. 22, 1984, pp. 98–114 and pp. 115–125, respectively.

Green, J., N. Wintfeld, P. Sharkey, and L. J. Passman, "The Importance of Severity of Illness in Assessing Hospital Mortality," *JAMA,* Vol. 263, No. 2, 1990, pp. 241–246.

Greenfield, S., H. U. Aronow, R. M. Elashoff, et al., "Flaws in Mortality Data: The Hazards of Ignoring Comorbid Disease," *JAMA,* Vol. 260, No. 15, 1988, pp. 2253–2255.

Hannan, E. L., J. F. O'Donnell, H. Kilburn, et al., "Investigation of the Relationship between Volume and Mortality for Surgical Procedures Performed in New York State Hospitals," *JAMA,* Vol. 262, No. 4, 1989, pp. 503–510.

Hausman, J., "Specification Tests in Econometrics," *Econometricia,* Vol. 46, No. 6, 1978, pp. 1251–1271.

Hebel, J. R., I. I. Kessler, K. Mabuchi, and R. J. McCarter, "Assessment of Hospital Performance by Use of Death Rates: A Recent Case History," *JAMA,* Vol. 248, No. 23, 1982, pp. 3131–3135.

Hsai, D. C., W. M. Krushat, A. B. Fagan, et al., "Accuracy of Diagnostic Coding for Medicare Patients under the Prospective-Payment System," *New Engl. J. Med.,* Vol. 318, 1988, pp. 352–355.

Jencks, S. F., J. Daley, D. Draper, et al., "Interpreting Hospital Mortality Data: The Role of Clinical Risk Adjustment," *JAMA,* Vol. 260, No. 24, 1988, pp. 3611–3624.

Jencks, S. F., D. K. Williams, and T. L. Kay, "Assessing Hospital-Associated Deaths from Discharge Data: The Role of Length of Stay and Comorbidities," *JAMA,* Vol. 260, No. 15, 1988, pp. 2240–2246.

Kahn, K. L., L. V. Rubenstein, M. R. Chassin, et al., *Medical Record Abstraction Form and Guidelines for Assessing Quality of Care for Hospitalized Patients with Congestive Heart Failure,* The RAND Corporation, N-2798-HCFA, December 1988.

Kahn, K. L., W. H. Rogers, L. V. Rubenstein, et al., "Measuring Quality of Care with Explicit Process Criteria Pre- and Post-Implementation of the DRG-Based Prospective Payment System," *JAMA,* Vol. 264, No. 11, 1990a, pp. 1969–1973.

Kahn, K. L., L. V. Rubenstein, D. Draper, et al., "The Effect of the DRG-Based Prospective Payment System on Quality of Care for Hospitalized Medicare Patients: Introduction to the Series" *JAMA,* Vol. 264, No. 11, 1990b, pp. 1953–1955.

Keeler, E. B., K. L. Kahn, D. Draper, et al., "Changes in Sickness at Admission Following the Introduction of Prospective Payment," *JAMA,* Vol. 264, No. 11, 1990, pp. 1962–1968.

Kennedy, J. W., G. C. Kaiser, L. D. Fisher, et al., "Multivariate Discriminant Analysis of the Clinical and Angiographic Predictors of Operative Mortality from the Collaborative Study in Coronary Artery Surgery (CASS)," *Journal Thorac. Cardiovasc. Surg.,* Vol. 80, 1980, pp. 876–887.

Knaus, W. A., E. A. Draper, D. P. Wagner, and J. E. Zimmerman, "An Evaluation of Outcome from Intensive Care in Major Medical Centers," *Ann. Int. Med.,* Vol. 104, 1986, pp. 410–418.

Kosecoff, J., K. L. Kahn, L. V. Rubenstein, et al., *Medical Record Abstraction Form and Guidelines for Assessing Quality of Care for Hospitalized Patients with Acute Myocardial Infarction,* The RAND Corporation, N-2799-HCFA, December 1988.

Lubitz, J., G. Riley, and M. Newton, "Outcomes of Surgery among the Medicare Aged: Mortality after Surgery," *Health Care Financing Review,* Vol. 6, 1985, pp. 103–115.

Luft, H. S., and S. S. Hunt, "Evaluating Hospital Quality Through Outcome Statistics," *JAMA,* Vol. 255, 1986, pp. 2780–2784.

Luft, H. S., J. P. Bunker, and A. C. Enthoven, "Should Operations Be Regionalized? The Empirical Relation between Surgical Volume and Mortality," *New Engl. J. Med.,* Vol. 301, 1979, pp. 1364–1369.

Manski, C. F., and S. R. Lerman, "The Estimation of Choice Probabilities from Choice Based Samples," *Econometrica,* November 1977.

Moses, L. E., "The Evaluation of Hospital Death Rates," *JAMA,* Vol. 255, No. 20, 1986, p. 2801.

Moses, L., and F. Mosteller, "Institutional Differences in Postoperative Death Rates: Commentary on Some of the Findings of the National Halothane Study," *JAMA,* Vol. 203, 1968, pp. 492–494.

Park, R. E., R. H. Brook, J. Kosecoff, et al., "Explaining Variations in Hospital Death Rates: Randomness, Severity of Illness, Quality of Care," *JAMA,* Vol. 264, No. 4, 1990, pp. 484–490.

Pollack, M. M., U. E. Ruttmann, P. R. Getson, and Members of the Multi-Institutional Study Group, "Accurate Prediction of the

Outcome of Pediatric Intensive Care: A New Quantitative Method," *New Engl. J. Med.,* Vol. 316, No. 3, 1987, pp. 134–139.

Riley, G., and J. Lubitz, "Outcomes of Surgery among the Medicare Aged: Surgical Volume and Mortality," *Health Care Financing Review,* Vol. 7, 1985, pp. 37–47.

Shortell, S. M., and J. P. LoGerfo, "Hospital Medical Staff Organization and Quality of Care: Results for Myocardial Infarction and Appendectomy," *Med. Care,* Vol. 19, 1981, pp. 1041–1055.

Showstack, J. A., K. E. Rosenfeld, D. W. Garnick, H. S. Luft, R. W. Schaffarzick, and J. Fowles, "Association of Volume with Outcome of Coronary Artery Bypass Graft Surgery: Scheduled vs Nonscheduled Operations" *JAMA,* Vol. 257, No. 6, 1987, pp. 785–789.

Sloan, F. A., J. M. Perrin, and J. Valvona, "In-Hospital Mortality of Surgical Patients: Is There an Empiric Basis for Standard Setting?" *Surgery,* Vol. 99, No. 4, 1986, pp. 446–453.

Stanford Center for Health Care Research, "Comparison of Hospitals with Regard to Outcomes of Surgery," *Health Services Research,* Vol. 11, 1976, pp. 112–127.

Steinbrook, R., "Hospital Death Rates of Medicare Patients Released," *The Los Angeles Times,* Part II, April 25, 1987, p. 1.

Sullivan, L. W., and L. B. Hays, *Medicare Hospital Mortality Information, 1986, 1987, 1988,* Health Care Financing Administration, publication number 00701, U.S. Department of Health and Human Services, 1989.

Wagner, D. P., W. A. Knaus, and E. A. Draper, "The Case for Adjusting Hospital Death Rates for Severity of Illness," *Health Affairs,* Vol. 5, No. 2, 1986, pp. 148–153.